Discretizing Unobserved Heterogeneity:

Approximate Clustering Methods for Dimension Reduction*

Stéphane Bonhomme† Thibaut Lamadon‡ Elena Manresa§

June 2016

INCOMPLETE DRAFT

Abstract

We develop two-step and iterative panel data estimators based on a discretization of unobserved heterogeneity. We view discrete estimators as approximations, and study their properties in environments where population heterogeneity is individual-specific and unrestricted, letting the number of types grow with the sample size. Bias reduction methods can improve the performance of discrete estimators. We also show that discrete estimation may strictly dominate fixed-effects approaches when unobservables are high-dimensional, provided their *underlying* dimension is low. We study two applications: a structural dynamic discrete choice model of migration, and a model of wage determination with worker and firm heterogeneity. These applications to settings with continuous heterogeneity suggest computational and statistical advantages of the discrete methods that we advocate.

1 Introduction

In nonlinear panel data models, fixed-effects approaches are conceptually attractive as they do not require restricting the form of unobserved heterogeneity. Recent theoretical developments in the literature include general treatments of asymptotic properties as both dimensions of the

*We thank Anna Simoni, Tim Christensen, Chris Hansen, Arthur Lewbel, Philippe Rigollet, and seminar audiences at various places for comments.

†University of Chicago, sbonhomme@uchicago.edu

‡University of Chicago, lamadon@uchicago.edu

§Massachusetts Institute of Technology, the Sloan School of Management, emanresa@mit.edu

panel increase, and methods for bias reduction and inference. See among others Arellano and Honoré (2001), Hahn and Newey (2004), and Arellano and Hahn (2007).

At the same time, fixed-effects approaches to nonlinear panel data estimation have not yet found wide applicability in empirical work. A practical challenge for these methods is that, in particular for computational reasons, it may be difficult or infeasible to allow for as many fixed-effects parameters as individual units in estimation.

Discrete approaches to unobserved heterogeneity potentially offer tractable alternatives. As an example, in the literature on structural dynamic discrete choice models, following Keane and Wolpin (1997) numerous papers have modeled individual heterogeneity as a small number of unobserved types. Discreteness is appealing in this context, as it leads to a finite number of unobserved state variables, and as it limits the number of parameters to estimate.

The analysis of discrete unobserved heterogeneity estimators was initially done from a random-effects perspective, under functional form and/or independence assumptions on unobservables and how they relate to observed covariates. Heckman and Singer (1984)'s analysis of single-spell duration models provides a seminal example of this approach. There is also a large literature on parametric and semi-parametric mixture models in statistics and econometrics.

A more recent literature aims at estimating the discrete individual-specific unobserved types directly, similarly as in fixed-effects approaches. The idea is to maintain the assumption that the statistical population consists of a finite number of types, while avoiding making further assumptions about unobservables. *Grouped fixed-effects* methods (Hahn and Moon, 2010, Bonhomme and Manresa, 2015) adapt and extend clustering techniques developed in machine learning and computer science to settings of interest in economic applications. This line of work has produced asymptotic "oracle" results under which the estimated type memberships converge to the true population types as both dimensions of the panel increase. Recent papers with similar properties in related settings are Lin and Ng (2012), Saggio (2012), Bai and Ando (2015), Su, Shi and Phillips (2015), and Vogt and Linton (2015).

In this paper we study discrete heterogeneity methods while being agnostic about the form of individual unobserved heterogeneity. In particular, we do not require the population to be discrete. We thus view the discrete modeling of heterogeneity as a dimension reduction device, instead of viewing discreteness as a substantive assumption about population unobservables. We study several grouped fixed-effects estimators, including novel moment-based two-step and iterative estimators which provide tractable ways of allowing for unobserved heterogeneity in nonlinear models.

Our asymptotic analysis stands in contrast with most of the literature on discrete methods, as we do not restrict the form of individual-specific unobservables. We show that grouped fixed-effects estimators generally suffer from an *approximation* bias that remains sizable unless the number of groups grows with the sample size. However, as the number of groups increases, estimating group membership becomes harder, and we show that this gives rise to an *incidental parameter* bias which has a similar order of magnitude as for conventional fixed-effects estimators. The central role played by group misclassification in this setting contrasts with asymptotic results derived in the previous literature on the topic. In addition, we show that in some circumstances our theory better reflects the finite-sample performance of grouped fixed-effects estimators even in discrete data generating processes, when the conditions for consistent "oracle" estimation are violated.

Our characterizations motivate the use of bias reduction and inference methods from the literature on fixed-effects nonlinear panel data estimation. Specifically, we use the split-panel jackknife method of Dhaene and Jochmans (2015) to reduce bias. Our results provide guarantees on the reliability of dimension reduction: for example, with scalar fixed-effects they apply as soon as the number of groups is of a similar or larger order of magnitude than the square root of the number of time periods.

In the second part of the theory, we explore what can be learned under restrictions on the underlying dimensionality of unobservables. Various economic settings feature multi-dimensional unobserved heterogeneity. In a structural model, unobserved types may drive ability, costs, or preferences, for example. Models with time-varying heterogeneity may feature as many unobservables as time periods. Under no further restrictions, for a moderate number of groups, the approximation bias is typically very large. However we show that, when the *underlying dimension* of heterogeneity is low, the bias typically decreases much faster as the number of groups increases. Examples of such settings are linear or nonlinear factor models with a small number of factors. We derive asymptotic rates of convergence which demonstrate that grouped structures are able to efficiently exploit commonalities between different dimensions of heterogeneity, in situations where fixed-effects estimators are not well-suited.[1]

To illustrate our results, we use grouped fixed-effects methods in two different settings. First, we consider structural dynamic discrete choice models. In these models, our two-step and iterative methods could provide useful alternatives to existing approaches in settings where time-series information is sufficiently informative.[2] We set up a simulation exercise based on

[1]The statistics literature on manifold learning and nonlinear principal components aims at learning about, and exploiting, the low intrinsic dimension of large dimensional data; see for example Hastie and Stuetzle (1989), Levina and Bickel (2004), and Raginsky and Lazebnik (2005).

[2]In particular, our methods can allow relying on sequential techniques for estimation, in a related but different

estimates from a simple dynamic model of location choice in the spirit of Kennan and Walker (2011), estimated on NLSY data. In a data generating process with continuous heterogeneity, we assess the magnitude of the biases of our discrete estimators and the usefulness of bias reduction.

Second, we revisit the estimation of workers' and firms' contributions to log-wage dispersion using matched employer-employee data. Focusing on a short panel version of the additive model of Abowd, Kramarz and Margolis (1999), we document through simulations calibrated to Swedish administrative data the magnitude of biases of one of the grouped fixed-effects estimators introduced in Bonhomme, Lamadon and Manresa (2015).

Our analysis borrows from previous work on *kmeans* clustering and vector quantization; see among others Gersho and Gray (1992), Gray and Neuhoff (1998), Graf and Luschgy (2000, 2002), Linder (2002), and Levrard (2015), as well as the seminal analysis of kmeans by Pollard (1981, 1982a, 1982b). There has been little work studying properties of discrete estimators as the sample size tends to infinity together with the number of groups. Important exceptions are Bester and Hansen (2016), who focus on a setup with known groups, and Gao, Lu and Zhou (2015) and Choi and Ohlede (2014), who derive results on stochastic blockmodels in networks.[3]

The outline of the paper is as follows. The first half focuses on the theory. We start in Section 2 by introducing two-step grouped fixed-effects estimators. We study their asymptotic properties in Section 3 and analyze iterative estimators in Section 4. In Section 5 we characterize some properties of grouped fixed-effects when unobservables are high-dimensional but their underlying dimension may be low. We then turn to the two applications in Sections 6 and 7. Lastly we conclude in Section 8.

2 Two-step grouped fixed-effects

We consider semi-parametric panel data models of the form $f(Y_i \,|\, X_i, \alpha_i, \theta)$, where θ are common parameters and α_i are q-dimensional individual effects. Using panel data notation we let $Y_i = (Y_{i1}, ..., Y_{iT})$ and $X_i = (X_{i1}', ..., X_{iT}')'$. We let $\ell_i(\alpha_i, \theta) = \ln f(Y_i \,|\, X_i, \alpha_i, \theta)/T$ be the average individual log-likelihood. The conditional distribution of α_i given X_i is left fully unrestricted. We are interested in estimating the vector of common parameters θ, as well as average effects depending on the α_i.

way to Arcidiacono and Jones (2003) and Arcidiacono and Miller (2011).

[3]Previous statistical analyses of stochastic blockmodels were done under discrete heterogeneity in the population; see for example Bickel and Chen (2009).

The estimation steps in our two-step grouped fixed-effects method are as follows. Let $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$ be individual-specific moments, which are informative about α_i in a sense to be made precise in the next section. In the first step, we estimate a partition of individual units, $\{\widehat{k}_i\}$, by finding the best grouped approximation to the moments $\{h_i\}$ based on K groups; that is:

$$\left(\widehat{\varphi}, \widehat{k}_1, ..., \widehat{k}_N\right) = \underset{(\varphi, k_1, ..., k_N)}{\operatorname{argmin}} \sum_{i=1}^{N} \left\| h_i - \varphi(k_i) \right\|^2, \tag{1}$$

where $\|\cdot\|$ denotes the Euclidean norm, $\{k_i\} \in \{1, ..., K\}^N$ are partitions of $\{1, ..., N\}$ into at most K groups, and $\varphi = (\varphi(1)', ..., \varphi(K)')'$ are $K \times 1$ vectors.

In the second step we maximize the log-likelihood function with respect to common parameters and group-specific individual effects, where the groups are given by the $\widehat{k}_i$ estimated in the first step; that is:

$$\left(\widehat{\theta}, \widehat{\alpha}\right) = \underset{(\theta, \alpha)}{\operatorname{argmax}} \sum_{i=1}^{N} \ell_i\left(\alpha\left(\widehat{k}_i\right), \theta\right), \tag{2}$$

where the maximization is with respect to θ and $\alpha = (\alpha(1)', ..., \alpha(K)')'$.

The optimization problem in (1) is referred to as *kmeans* in machine learning and computer science. It applies to general moment vectors h_i. In (1) the minimum is taken with respect to all possible partitions, in addition to values $\varphi(1), ..., \varphi(K)$. Computing a global minimum may be challenging, yet fast and stable heuristic algorithms exist, such as iterative descent, genetic algorithms or variable neighborhood search.[4] In this paper, consistently with most of the statistical literature on classification estimators (e.g., Pollard, 1981, 1982a) we will focus on the properties of the global minimum in (1).

Note that the optimization problem in (2) is of a lower dimension compared to fixed-effects maximum likelihood, as one only needs to estimate K values $\alpha(1), ..., \alpha(K)$ instead of N individual-specific parameters $\alpha_1, ..., \alpha_N$.

Example 1: dynamic discrete choice model. A prototypical structural dynamic discrete choice model features the following elements (see for example Aguirregabiria and Mira, 2010): choices $j_{it} \in \{1, ..., J\}$, payoff variables Y_{it}, and observed and unobserved state variables X_{it} and α_i, respectively. As an example, in the location choice model of Section 6, j_{it} is location at time t, and log-wages Y_{it} depend on latent location-specific returns $\alpha_i(j_{it})$. The individual

[4] See Steinley (2001) and Bonhomme and Manresa (2015) for references.

log-likelihood function conditional on initial choices and state variables typically takes the form:

$$\ell_i(\alpha_i, \theta) \quad = \quad \sum_{t=2}^{T} \ln f\left(j_{it} \,|\, X_{it}, \alpha_i, \theta\right) + \ln f\left(X_{it} \,|\, j_{i,t-1}, X_{i,t-1}, \alpha_i, \theta\right) + \ln f\left(Y_{it} \,|\, j_{it}, X_{it}, \alpha_i, \theta\right). \tag{3}$$

Computing choice probabilities $f\left(j_{it} \,|\, X_{it}, \alpha_i, \theta\right)$ in (3) requires solving the dynamic optimization problem, which can be demanding. In two-step grouped fixed-effects, one estimates a partition $\{\widehat{k}_i\}$ in a first step that does not require solving the model. In the second step, the partition $\{\widehat{k}_i\}$ is taken as given and the log-likelihood in (3) is maximized with respect to θ and type-specific parameters $\alpha(k)$. This may reduce the computational burden compared both to fixed-effects maximum likelihood, and to random-effects mixture approaches which are commonly based on iterative algorithms.

As h_i vectors to be used in the first step, one may take moments of payoff variables, observed state variables, and choices. One may also use individual-specific conditional choice probabilities $f_i(j_{it} = j \,|\, X_{it} = x)$, possibly based on a coarsening of covariates X_{it}. In the application in Section 6 we use means of log-wages in a first step, while relying on a likelihood-based iteration which exploits the full model's structure, hence also using information on choices, to improve estimation accuracy.

Example 2: linear regression. We will use a simple regression example to illustrate our assumptions and results. Consider the following model for a scalar outcome:

$$Y_{it} = \rho Y_{i,t-1} + X_{it}'\beta + \eta_i + \varepsilon_{it}, \tag{4}$$

where $|\rho| < 1$. A two-step grouped fixed-effects estimator in this model can be based on the moment vector $h_i = (\overline{Y}_i, \overline{X}_i')'$. The second step then consists in regressing Y_{it} on $Y_{i,t-1}$, X_{it}, and group indicators.

3 Asymptotic properties of two-step estimators

In this section we study the two estimation steps in turn.

3.1 First step

Our first result is to derive a rate of convergence for the kmeans estimator $\widehat{\varphi}(\widehat{k}_i)$ in (1). Let q and $r \geq q$ denote the dimensions of α_i and h_i, respectively. In this section and the next q

and r are kept fixed as N, T tend to infinity. Throughout we use 0 subscripts to indicate true parameter values. We make the following assumptions.

Assumption 1. *(moments, first step)*

(i) As N, T tend to infinity, $\frac{1}{N}\sum_{i=1}^N \|h_i - \varphi(\alpha_{i0})\|^2 = O_p\left(\frac{1}{T}\right)$.

(ii) φ is Lipschitz continuous.

The first part in Assumption 1 implies that $h_i = \frac{1}{T}\sum_{t=1}^T h(Y_{it}, X_{it})$ is $\sqrt{T}$-consistent for $\varphi(\alpha_{i0})$. This will be true under weak conditions on serial dependence of $\varepsilon_{it} = h(Y_{it}, X_{it}) - \varphi(\alpha_{i0})$, such as suitable α-mixing conditions, which are commonly made when studying asymptotic properties of fixed-effects panel data estimators. The main assumption is that the probability limit of h_i is only a function of α_{i0}. As h_i is a moment of the joint distribution (Y_i, X_i), its probability limit will generally depend on the individual effects that enter the conditional distribution of Y_i given X_i, but also on the individual effects that enter the marginal distribution of covariates X_i. As we now illustrate in the regression example, Assumption 1 typically requires defining α_{i0} as containing both types of individual effects.

Example 2 (continued). Write (4) in compact form as $Y_{it} = W_{it}'\theta_0 + \eta_{i0} + \varepsilon_{it}$, where $W_{it} = (Y_{i,t-1}, X_{it}')'$ and $\theta_0 = (\rho_0, \beta_0')'$. Let $\mu_{i0} = \text{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^T X_{it}$, and let $\alpha_{i0} = (\eta_{i0}, \mu_{i0}')'$. The vector α_{i0} contains two types of individual effects: those that enter the outcome distribution conditional on covariates (that is, η_{i0}), and those that enter the distribution of covariates (that is, μ_{i0}). Note that there could be additional heterogeneity in the variance of h_i, for example, which needs not be included in α_{i0}. For the moment vector $h_i = (\overline{Y}_i, \overline{X}_i')'$, we then have under standard conditions: $\text{plim}_{T\to\infty}\, h_i = \left(\frac{\eta_{i0}+\mu_{i0}'\beta_0}{1-\rho_0}, \mu_{i0}'\right)' = \varphi(\alpha_{i0})$.

Let us define the following quantity, which we refer to as the *approximation bias* of α_{i0}:

$$B_\alpha(K) = \min_{(\alpha, \{k_i\})} \frac{1}{N}\sum_{i=1}^N \|\alpha_{i0} - \alpha(k_i)\|^2,$$

where, similarly as in (1), the minimum is taken with respect to all $\{k_i\}$ and $\alpha(k)$. The term $B_\alpha(K)$ represents the approximation error one would make if one were to discretize the population unobservables α_{i0} directly. It is typically a decreasing function of K. Below we review existing results about the magnitude of $B_\alpha(K)$ in various settings.

We have the following characterization of the rate of convergence of $\widehat{\varphi}(\widehat{k}_i)$. In the asymptotic we let $T = T_N$ and $K = K_N$ tend to infinity jointly with N. All proofs are in the appendix.

Theorem 1.

Let Assumption 1 hold. Then, as N, T, K tend to infinity:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\varphi}(\widehat{k}_i)-\varphi(\alpha_{i0})\right\|^2 = O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right).$$

Theorem 1 provides an upper bound on the rate of convergence of the discrete estimator $\widehat{\varphi}(\widehat{k}_i)$ of $\varphi(\alpha_{i0})$. This bound is larger than the rate of the fixed-effects estimator $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$, which is $O_p(1/T)$ under the same conditions.

Approximation bias. The rate in Theorem 1 depends on the approximation bias $B_\alpha(K)$, which is a key quantity in the analysis of the estimators we consider in this paper. This quantity has been extensively studied in the literature on vector quantization, where it is referred to as the empirical quantization error.[5]

The approximation bias $B_\alpha(K)$ is closely related to the dimension of unobserved heterogeneity. Graf and Luschgy (2002) provide explicit characterizations in the case where α_{i0} has compact support with a nonsingular probability distribution. As N, K tend to infinity, their Theorem 5.3 establishes that $B_\alpha(K) = O_p(K^{-\frac{2}{q}})$. This implies that $B_\alpha(K) = O_p(K^{-2})$ when α_{i0} is one-dimensional, and $B_\alpha(K) = O_p(K^{-1})$ when α_{i0} is two-dimensional, for example. In models with conditioning covariates such as Example 2, in the presence of a large number of heterogeneous covariates, moment-based two-step methods may suffer from a large approximation bias. This motivates considering alternative estimators based on iterative approaches, see Section 4.

The quality of approximation of the discretization depends on the underlying dimensionality of the heterogeneity, as opposed to its number of components. To see this, note that, if $\alpha_{i0} = a(\xi_{i0})$, where ξ_{i0} has dimension $d \leq q$, and $a(\cdot)$ is a Lipschitz continuous function, then $B_\alpha(K) = O_p(B_\xi(K))$.[6] Likewise, as φ is Lipschitz by the first part of Assumption 1, $B_\varphi(K) = O_p(B_\alpha(K))$. As a result, for moderate K the approximation bias may still be relatively small when α_{i0} has a large number of dimensions, provided these dimensions are linked to each other in some way

[5] Empirical quantization errors can be mapped to covering numbers commonly used in empirical process theory. Specifically, it can be shown that if the ϵ-covering number (for the Euclidean norm) of the set $\{\alpha_1, ..., \alpha_N\}$ is such that $\mathcal{N}(\epsilon, \{\alpha_i\}, \|\cdot\|) \geq K$, then $B_\alpha(K) \leq \epsilon^2$.

[6] This is a direct consequence of the fact that, if $\|a(\xi') - a(\xi)\| \leq \mu\|\xi' - \xi\|$ for all (ξ, ξ'), then:

$$\min_{(\alpha,\{k_i\})}\frac{1}{N}\sum_{i=1}^{N}\|\alpha_{i0}-\alpha(k_i)\|^2 \leq \mu^2 \cdot \min_{(\alpha,\{k_i\})}\frac{1}{N}\sum_{i=1}^{N}\|\xi_{i0}-\xi(k_i)\|^2.$$

and the underlying dimension of α_{i0} is low. In those cases, discretizing the data jointly as in (1) may allow exploiting the presence of such a low dimension in the data, as opposed to discretizing each component of h_i separately. We will explore this type of situations, examples of which include nonlinear factor models and models with time-varying heterogeneity, in Section 5.

The next result establishes a bound on the difference between the sample mean h_i and the kmeans estimator $\widehat{\varphi}(\widehat{k}_i)$, under some conditions on approximation biases.

Corollary 1. *Let $\varepsilon_i = \frac{1}{T}\sum_{t=1}^T h(Y_{it}, X_{it}) - \varphi(\alpha_{i0}) = h_i - \varphi(\alpha_{i0})$. Suppose that φ is Lipschitz continuous. Suppose that as N, K tend to infinity $B_\alpha(K) = O_p(K^{-\frac{2}{d}})$, and as N, T, K tend to infinity $B_{\sqrt{T}\varepsilon}(K) = O_p(K^{-\frac{2}{r}})$. Then, as N, T, K tend to infinity such that $TK^{-\frac{2}{d}}$ tends to zero:*

$$\frac{1}{N}\sum_{i=1}^N \left\| h_i - \widehat{\varphi}(\widehat{k}_i)\right\|^2 = O_p\left[\frac{1}{T}\left(\frac{T}{K^{\frac{2}{d}}}\right)^{\frac{d}{d+r}}\right].$$

Under the conditions of Corollary 1, the fixed-effects estimator $h_i = \frac{1}{T}\sum_{t=1}^T h(Y_{it}, X_{it})$ and the grouped fixed-effects estimator $\widehat{\varphi}(\widehat{k}_i)$ have the same rate $C/T + o_p(1/T)$, for a constant $C > 0$, provided $TK^{-\frac{2}{d}}$ tends to zero. For example, when α_{i0} is scalar this will happen as soon as TK^{-2} tends to zero. This result thus refines the upper bound on the rate of convergence obtained in Theorem 1. As r increases for d fixed, the rate in Corollary 1 becomes arbitrarily close to $O_p(1/T)$. Consistently with this observation, in Section 5 we will see that in high-dimensional settings grouped fixed-effects may strictly dominate fixed-effects when the underlying dimension of heterogeneity (that is, d) is low. Finally, note that here we do not provide primitive assumptions for the conditions of Corollary 1, as this appears difficult.[7]

3.2 Second step

We now turn to the second estimation step. We make the following assumptions.

Assumption 2. *(regularity)*

$\ell_i(\alpha_i, \theta)$ is three times differentiable in both its arguments, with bounded derivatives. Observations are i.i.d. across individual units, and weakly dependent over time. θ_0 is identified. In addition, the minimum eigenvalue of $\mathbb{E}\left(-\frac{\partial^2 \ell_i(\alpha_i,\theta)}{\partial \alpha_i \partial \alpha_i'}\right)$ is asymptotically bounded away from zero uniformly in (α_i, θ).

[7] While results on empirical quantization errors have been derived in the large-N limit under general conditions, see for example Theorem 6.2 in Graf and Luschgy (2000), rates as N and K tend to infinity jointly are so far limited to distributions with compact support; see the remark on p. 875 in Graf and Luschgy (2002).

The main condition in Assumption 2 is the concavity of the likelihood function with respect to α_i. At this stage it is not clear whether this condition is needed, or if it is an artifact of our method of proof. The assumptions on dependence are standard. They could be relaxed to allow for weak dependence across individual units.

Assumption 3. *(moments, second step)*

(i) There exists a Lipschitz continuous function ψ such that $\alpha_{i0} = \psi(\varphi(\alpha_{i0}))$.

(ii) The probability limits of the following quantities do not depend on i beyond α_{i0}:

$$\frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\theta\partial\alpha_i'}, \quad \frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\alpha_i\partial\alpha_i'}, \quad \text{and} \quad \frac{\partial \ell_i(\alpha,\theta)}{\partial\alpha_i} \text{ for all } (\alpha,\theta).$$

The first part of Assumption 3 requires the individual moment $h_i = \frac{1}{T}\sum_{t=1}^T h(Y_{it}, X_{it})$ to be informative about α_{i0}, in the sense that $\text{plim}_{T\to\infty}\, h_i = \varphi(\alpha_{i0})$, and $\alpha_{i0} = \psi(\varphi(\alpha_{i0}))$. The second part requires the probability limits of some particular moments to only be functions of α_{i0}. Neither φ nor ψ need to be known to the econometrician.

Example 2 (continued). In Example 2, letting $\psi(\varphi_{i1},\varphi_{i2}) = ((1-\rho_0)\varphi_{i1} - \varphi_{i2}'\beta_0, \varphi_{i2}')'$, we have $\alpha_{i0} = \psi(\varphi(\alpha_{i0}))$. Using a grouped fixed-effects estimator based on a Gaussian pseudo-likelihood, Assumptions 2 and 3 can then be verified under standard conditions on error terms and covariates.

The next result provides asymptotic properties on two-step grouped fixed-effects estimators.

Theorem 2.

Let Assumptions 1, 2 and 3 hold. Let:

$$H = \lim_{N,T\to\infty} -\frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[\frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\theta\partial\theta'} - \frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\theta\partial\alpha_i'}\left(\frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\alpha_i\partial\alpha_i'}\right)^{-1}\frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial\alpha_i\partial\theta'}\right].$$

Then, as N, T, K tend to infinity:

$$\widehat{\theta} = \theta_0 + H^{-1}\frac{1}{N}\sum_{i=1}^N \frac{\partial \ell_i(\alpha_{i0},\theta_0)}{\partial\theta} + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) + o_p\left(\frac{1}{\sqrt{NT}}\right), \tag{5}$$

and:

$$\frac{1}{N}\sum_{i=1}^N \left\|\widehat{\alpha}(\widehat{k}_i) - \alpha_{i0}\right\|^2 = O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right). \tag{6}$$

Theorem 2 implies that, when the approximation bias $B_\alpha(K)$ is of a similar or lower order of magnitude as $1/T$, the asymptotic distribution of a two-step grouped fixed-effects estimator has a similar structure as a conventional fixed-effects estimator, as studied by Hahn and Newey (2004) among others. Like fixed-effects, grouped fixed-effects estimators suffer in general from an $O_p(1/T)$ bias term.

Bias correction. The link to fixed-effects motivates adapting existing bias reduction techniques to grouped fixed-effects estimation. A variety of methods have been developed in the nonlinear panel data literature; see Arellano and Hahn (2007) for a review. Here we consider the split-panel jackknife method of Dhaene and Jochmans (2015). The method works as follows.

We first estimate $\widehat{\theta}$ using two-step grouped fixed-effects on the full sample. Then, we estimate $\widehat{\theta}_1$ and $\widehat{\theta}_2$ on the first $T/2$ periods (considering T even for simplicity) and the last $T/2$ periods, respectively. The bias-reduced estimator is then:

$$\widehat{\theta}^{BR} = 2\widehat{\theta} - \frac{\widehat{\theta}_1 + \widehat{\theta}_2}{2}.$$

In practice, we recommend keeping the product $TB_\alpha(K)$ (asymptotically) constant in all three samples. For example, when α_i is scalar and $B_\alpha(K)$ is of the order of $1/K^2$, we recommend taking the integer part of $K/\sqrt{2}$ in each half-panel. The half-panel jackknife method requires stationary panel data, however it can allow for serial correlation and dynamics.

To derive the asymptotic distribution of $\widehat{\theta}^{BR}$, let us suppose that, as N, T, K tend to infinity, the $O_p(1/T)$ term on the right-hand side of (5) takes the form $C/T + o_p(1/T)$, for some constant $C > 0$. From Theorem 2, the bias-reduced grouped fixed-effects estimator then has the following distribution as N, T, K tend to infinity such that N/T and $TB_\alpha(K)$ tend to constants:

$$\sqrt{NT}\left(\widehat{\theta}^{BR} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, \Omega). \tag{7}$$

In (7), Ω coincides with the asymptotic variance of the two-step grouped fixed-effects estimator, which itself coincides with that of the fixed-effects estimator; that is:

$$\Omega = H^{-1}\mathbb{E}\left[\frac{\partial \ell_i(\alpha_{i0}, \theta_0)}{\partial \theta}\frac{\partial \ell_i(\alpha_{i0}, \theta_0)}{\partial \theta'}\right] H^{-1}.$$

The asymptotic variance can be consistently estimated using several methods, for example using a HAC formula clustered at the individual level, replacing α_{i0} and θ_0 by their grouped fixed-effects estimates $\widehat{\alpha}(\widehat{k}_i)$ and $\widehat{\theta}$ (or by their bias-reduced grouped fixed-effects estimates).

Although Theorem 2 does not provide direct guidance on the choice of the number of groups K, the result can be used to justify dimension reduction based on a given K. As an example, consider a case when α_{i0} is scalar and $B_\alpha(K) = O_p(1/K^2)$. In this case Theorem 2 justifies conducting inference on θ_0 based on a split-panel jackknifed estimator and equation (7), provided K is of a similar or larger order of magnitude as $\sqrt{T}$. In situations where $\sqrt{T}$ is small relative to N and the likelihood function is hard to evaluate, this can represent a substantial decrease in computational cost compared to fixed-effects estimation.

Comparison with results under discrete heterogeneity. It is useful to compare Theorem 2, obtained in an environment where the population is not grouped, to existing results on the performance of grouped fixed-effects estimators in discrete population settings. When the population consists of K^* groups, where K^* is known,[8] Hahn and Moon (2010) and Bonhomme and Manresa (2015) provide conditions under which estimated group membership $\widehat{k}_i$ tends in probability to the population group membership k_i^*, up to arbitrary labeling of the groups. Their conditions imply that the probability of misclassifying at least one individual unit tends to zero as N, T tend to infinity and N/T^δ tends to zero for some $\delta > 0$. In this asymptotic the grouped fixed-effects estimators are not affected by incidental parameter bias. In other words, the asymptotic distribution of $\widehat{\theta}$ is not affected by the fact that group membership has been estimated.

These existing characterizations of grouped fixed-effects in discrete populations contrast with Theorem 2. Indeed, the latter implies that, in a world with possibly non-discrete heterogeneity, classification noise needs to be taken into account in the properties of the estimators. As the number of groups increases to control approximation bias, group membership estimates $\widehat{k}_i$ become increasingly noisy as groups become harder to distinguish. This causes the presence of an incidental parameter bias, the magnitude of which leads to the same $1/T$ bias as conventional fixed-effects estimators. Providing valid inference based on grouped fixed-effects thus requires working in a very different asymptotic framework. In fact, as our simulations below show, the characterizations in this section may also provide more accurate inference in situations where the population is discrete, but the conditions for applying existing "oracle" results from the literature are violated.

Average effects. Average effects are also often of interest in economic settings. Let $m_i(\alpha_i, \theta)$ be a shorthand for $\frac{1}{T}\sum_{t=1}^{T} m(X_{it}, \alpha_i, \theta)$. A grouped fixed-effects estimator of the population

[8] Under suitable conditions K^* can be consistently estimated using information criteria or sequential tests.

average $M_0=\frac{1}{N}\sum_{i=1}^N m_i\left(\alpha_{i0},\theta_0\right)$ is:

$$\widehat{M}=\frac{1}{N}\sum_{i=1}^N m_i\left(\widehat{\alpha}(\widehat{k}_i),\widehat{\theta}\right).$$

Assumption 4. *(average effects)*

$m_i(\alpha_i,\theta)$ is twice differentiable with respect to α_i and θ, with bounded derivatives. Moreover, the probability limit of $\frac{\partial m_i(\alpha_{i0},\theta_0)}{\partial \alpha_i}$ does not depend on i beyond α_{i0}.

Corollary 2. *Let the assumptions of Theorem 2 hold. Let Assumption 4 hold. Let $a_i=\frac{\partial m_i(\alpha_{i0},\theta_0)}{\partial \alpha_i'}\left(-\frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial \alpha_i\partial\alpha_i'}\right)^{-1}$, and $b=\frac{1}{N}\sum_{i=1}^N\left(a_i\frac{\partial^2 \ell_i(\alpha_{i0},\theta_0)}{\partial \alpha_i\partial\theta'}+\frac{\partial m_i(\alpha_{i0},\theta_0)}{\partial\theta'}\right)H^{-1}$, where H is defined as in Theorem 2.*

Then, as N,T,K tend to infinity:

$$\begin{aligned}\widehat{M} &= M_0+\frac{1}{N}\sum_{i=1}^N a_i\frac{\partial \ell_i(\alpha_{i0},\theta_0)}{\partial\alpha_i}+b\cdot\frac{1}{N}\sum_{i=1}^N\frac{\partial \ell_i(\alpha_{i0},\theta_0)}{\partial\theta}\\ &\quad +O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right)+o_p\left(\frac{1}{\sqrt{NT}}\right).\end{aligned}$$

The same bias-reduction and inference techniques that we use for common parameters θ_0 can be used when estimating average effects M_0.

4 Iterative and one-step estimators

Two-step grouped fixed-effects relies on a suitable choice of moments to be discretized in a first step. In applications, it may be useful to exploit the model's structure to guide this choice. This can be done by iterating the grouped fixed-effects method, as we now describe.

Starting with two-step estimates $\widehat{\theta}$ and $\widehat{\alpha}$ from (2), a new partition of individual units can be computed according to the following classification rule:

$$\widehat{k}_i^{(2)}=\underset{k\in\{1,...,K\}}{\operatorname{argmax}}\ \ell_i\left(\widehat{\alpha}\left(k\right),\widehat{\theta}\right),\qquad \text{for all } i=1,...,N. \tag{8}$$

Then, second-step estimates can be updated as:

$$\left(\widehat{\theta}^{(2)},\widehat{\alpha}^{(2)}\right)=\underset{(\theta,\alpha)}{\operatorname{argmax}}\ \sum_{i=1}^N \ell_i\left(\alpha\left(\widehat{k}_i^{(2)}\right),\theta\right). \tag{9}$$

The method may be iterated.

We have the following result for $\widehat{\theta}^{(2)}$. Similar results hold for $\widehat{\alpha}^{(2)}(\widehat{k}_i^{(2)})$ and average effects, although we omit them for brevity.

Corollary 3. *Let the assumptions of Theorem 2 hold. Let H as in Theorem 2. Then, as N, T, K tend to infinity:*

$$\widehat{\theta}^{(2)} = \theta_0 + H^{-1} \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ell_i(\alpha_{i0}, \theta_0)}{\partial \theta} + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) + o_p\left(\frac{1}{\sqrt{NT}}\right).$$

Multi-step iterative estimators based on this strategy have a similar asymptotic behavior as the one characterized in Corollary 3.

One-step estimation. A related estimator is the one-step grouped fixed-effects estimator, which is defined as follows:

$$\left(\widehat{\theta}^{1step}, \widehat{\alpha}^{1step}, \{\widehat{k}_i^{1step}\}\right) = \underset{(\theta,\alpha,\{k_i\})}{\operatorname{argmax}} \sum_{i=1}^{N} \ell_i\left(\alpha\left(k_i\right), \theta\right), \tag{10}$$

where the maximum is taken with respect to all possible parameter values θ and α and all possible partitions $\{k_i\}$ of $\{1, ..., N\}$ into at most K groups. This corresponds to the classification maximum likelihood estimator (e.g., Bryant and Williamson, 1978) which was studied in Hahn and Moon (2010) and Bonhomme and Manresa (2015). Unlike moment-based estimation first steps (see (1)) the one-step estimator in (10) requires evaluating the likelihood function for every partition and parameter values considered. It may thus be substantially more computationally intensive than two-step or multi-step methods based on an initial choice of moments.

Note that the objective function in (10) of the one-step estimator is not smooth since the discrete classification depends on parameter values. In the case of kmeans, Pollard (1981, 1982a) derived asymptotic properties for fixed K and T as N tends to infinity. Deriving the properties of one-step estimators in (10) as N, T, K tend jointly to infinity is an interesting avenue for future work. We now return to Example 2 and characterize properties of two-step and one-step grouped fixed-effects estimators in more detail.

Example 2 (continued). By Theorem 2, under weak dependence conditions on error terms ε_{it}, the two-step estimators of ρ_0 and β_0 based on $h_i = (\overline{Y}_i, \overline{X}_i')'$ in model (4) have bias $O_p(1/T) + O_p(B_\alpha(K))$. Note that, as the dimension of X_{it} increases, $B_\alpha(K)$ decreases at a slower rate as a function of K.

In the appendix we derive the first-order bias term for the one-step estimator in model (4). Under normality, the bias takes a simple form that combines the bias from within-groups with a "between" component which tends to zero as the number of groups increases. The rate of convergence of the approximation bias is $1/K^2$ in this case, irrespective of the dimension of the vector of covariates. This reflects the fact that one-step estimation delivers model-based moments which can improve the performance of grouped fixed-effects.

Lastly, in model (4) one can consider other possibilities for estimation. In the appendix we present a "double grouped fixed-effects" estimator where we discretize all components of h_i and include the indicators of estimated groups additively in the second-step regression. This strategy can be used in linear or linear-index models. In the appendix we report simulation results for a probit model.

5 High-dimensional unobservables

In this section we study properties of grouped fixed-effects estimation in settings where unobservables are high-dimensional. We explore the usefulness of assumptions on the underlying dimensionality of unobservables. When unobservables are multi-dimensional, fixed-effects estimators may not be well-suited while grouped fixed-effects estimators may still perform well provided the underlying dimension of α_i is low.

5.1 Rate of convergence

We consider situations where q, the dimension of α_i, tends to infinity together with N, T, K. We focus on the first step of the two-step method. Let $h_i = h(Y_i, X_i)$ be an r-dimensional vector of data moments with $h_i = \varphi(\alpha_{i0}) + \varepsilon_i$. As an example, $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$ and $\varepsilon_i = \frac{1}{T}\sum_{t=1}^{T}\left(h(Y_{it}, X_{it}) - \varphi(\alpha_{i0})\right)$. The function φ maps $\mathbb{R}^q$ to $\mathbb{R}^r$. We make the following assumptions.

Assumption 5. *(tails)*

(i) $\mathbb{E}\left(\|\varepsilon_i\|^2\right) \leq \sigma_T^2$, *for* $i = 1, ..., N$.

(ii) The vector $\varepsilon = (\varepsilon_1', ..., \varepsilon_N')'$ *has sub-Gaussian tails; that is, there exists* $c_T > 0$ *such that:*

$$\mathbb{E}\left[\exp(\tau'\varepsilon)\right] \leq \exp(c_T \cdot \|\tau\|^2) \quad \textit{for all } \tau \in \mathbb{R}^{rN}.$$

The first part in Assumption 5 bounds the second moment of ε_i. The second part requires the tails of ε_i to be sub-Gaussian (e.g., Vershynin, 2010). For example, i.i.d. Gaussian random variables and i.i.d. bounded random variables are sub-Gaussian. More generally, this assumption allows for dependence across observations. For example, in the case where $\varepsilon \sim \mathcal{N}(0, \Sigma)$ Assumption 5 holds provided the maximal eigenvalue of Σ is bounded from above by $2c_T$. Taking c_T to be asymptotically constant, this allows for weak forms of dependence in ε_i, across both individual units and components of ε_i.[9] When ε_i are i.i.d. across units i, letting $\Sigma = \mathbb{E}\left(\varepsilon_i\varepsilon_i'\right)$ we can take σ_T^2 to be the trace of Σ, and $2c_T$ to be the maximal eigenvalue of Σ. Note that in this case $c_T \leq \sigma_T^2/2$.

Assumption 6. *(regularity)*

The ratio r/q tends to a positive constant as q diverges. The function φ is Lipschitz continuous with uniformly bounded Lipschitz coefficient.

Theorem 3.

Let Assumptions 5 and 6 hold. Then, as N, T, K, q tend to infinity:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\varphi}(\widehat{k}_i)-\varphi(\alpha_{i0})\right\|^2 = \min\left\{O_p\left(c_T\cdot\ln K\right)+O_p\left(c_T\cdot\frac{qK}{N}\right)+O_p\left(B_\alpha(K)\right),\right.$$
$$\left. O_p\left(\sigma_T^2\right)+O_p\left(B_\alpha(K)\right)\right\}. \tag{11}$$

An immediate corollary of Theorem 3 is that, if there exists a Lipschitz continuous function ψ such that $\alpha_{i0}=\psi(\varphi(\alpha_{i0}))$, then $\frac{1}{N}\sum_{i=1}^{N}\left\|\psi\left(\widehat{\varphi}(\widehat{k}_i)\right)-\alpha_{i0}\right\|^2$ has the same rate of convergence as in (11). This result may be used to derive properties of two-step estimators, similarly as in Theorem 2.

5.2 Examples

We now illustrate through two examples that grouped fixed-effects estimators may asymptotically dominate fixed-effects ones when the underlying dimensionality of α_i is low.

Vector-valued fixed-effects. As a first example where the dimension of α_i grows with the sample size, consider the model $Y_{it}=\alpha_{i0}+\varepsilon_{it}$, where Y_{it} has dimension q. Consider the moments $h_i=\overline{Y}_i$, with $r=q$. Suppose that Assumption 5 holds for $\varepsilon_i=\frac{1}{T}\sum_{t=1}^{T}\varepsilon_{it}$, with $\sigma_T^2=Cq/T+o(q/T)$ for some constant $C>0$, and $c_T=O(1/T)$. For example, the latter holds if ε_{it} are i.i.d. across units i, and $(\varepsilon_{i1}',...,\varepsilon_{iT}')'$ are Gaussian $(0,\Sigma)$, with the maximum

[9] For example, in the case where $r=T$ and $\varepsilon_i=(\varepsilon_{i1},...,\varepsilon_{iT})'$, this allows for dependence across individuals and over time. Related conditions have been used in the literature on large approximate factor models (Chamberlain and Rothschild, 1983, Bai and Ng, 2002).

eigenvalue of Σ being uniformly bounded. This requires serial dependence to be weak. Then Theorem 3 implies that, as N, T, K, q tend to infinity:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2 = O_p\left(\frac{\ln K}{T}\right)+O_p\left(\frac{qK}{NT}\right)+O_p\left(B_{\alpha}(K)\right),$$

where for example $B_{\alpha}(K)=O_p(qK^{-2})$ when the underlying dimension of α_{i0} is one. In contrast, the rate of the fixed-effects estimator is:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{Y}_i-\alpha_{i0}\right\|^2 = C\frac{q}{T}+o_p\left(\frac{q}{T}\right).$$

Hence, in this environment, grouped fixed-effects may improve over fixed-effects when $\ln K \leq q$.

Time-varying heterogeneity. As a second example, consider the following specification with scalar Y_{it} and time-varying unobserved heterogeneity: $Y_{it}=\alpha_{i0}(t)+\varepsilon_{it}$. In this case we consider the moments $h_i = Y_i$, with $r = q = T$. Suppose that Assumption 5 holds for $\varepsilon_i=(\varepsilon_{i1},...,\varepsilon_{iT})'$, with $\sigma_T^2=O(1)$ and $c_T=O(1)$. This will hold for example if $(\varepsilon_1',...,\varepsilon_N')'$ are Gaussian $(0,\Sigma)$, with the maximum eigenvalue of Σ uniformly bounded. Then Theorem 3 implies that, as N, T, K tend to infinity:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{\alpha}(\widehat{k}_i,t)-\alpha_{i0}(t)\right)^2 = O_p\left(\frac{\ln K}{T}\right)+O_p\left(\frac{K}{N}\right)+O_p\left(\frac{B_{\alpha}(K)}{T}\right). \tag{12}$$

In this case there is of course no fixed-effects counterpart to the grouped fixed-effects estimator.

The three components in the rate of convergence in (12) are intuitive. The K/N part reflects the fact that we are estimating KT parameters (that is, the $\alpha(k,t)$) using NT observations.[10] The $(\ln K)/T$ term is equal to the logarithm of the number of possible classifications of N individual units into K groups (that is, K^N) divided by the number NT of observations. The term reflects the presence of an incidental parameter bias due to noisy group classification, similarly as the $1/T$ term in Theorem 1.[11]

The third component of the rate in (12) is $B_{\alpha}(K)/T$. As in the case studied in Section 2 this term depends on the underlying dimensionality of $\alpha_{i0}(t)$. When no restrictions are made on $\alpha_{i0}(t)$ except bounded support, one can only bound $B_{\alpha}(K)/T$ by $K^{-\frac{2}{T}}$, which does not tend to zero unless K is extremely large relative to T.[12] Restricting the underlying dimension of $\alpha_{i0}(t)$

[10] Here we are estimating unrestricted group-time effects $\alpha(k,t)$. Under smoothness restrictions on $\alpha_{i0}(t)$ as a function of t the K/N term could be reduced.

[11] A related term arises in the study of stochastic blockmodels for networks by Gao, Lu and Zhou (2015).

[12] Note that $K^{-\frac{2}{T}}$ does not tend to zero asymptotically unless $(\ln K)/T$ tends to infinity.

is intuitive in order to separate its contribution from that of the latent ε_{it}. Examples with low underlying dimensionality are linear or nonlinear factor models of the form $\alpha_{i0}(t) = \alpha(\xi_{i0}, t)$, where α is Lipschitz continuous in its first argument and the factor loading ξ_{i0} has dimension d. In that case, under suitable conditions, the approximation bias is $B_\alpha(K)/T = O_p(K^{-\frac{2}{d}})$.

Lastly, the rate in Theorem 3 suggests ways of choosing K in practice. As an example, consider the time-varying heterogeneous case of (12), and suppose that $B_\alpha(K)/T = O_p\left(K^{-\frac{2}{d}}\right)$ and that N and T tend to infinity at the same rate. In this case, the number of groups minimizing the right-hand side of (12) is $K^* = N^{\frac{d}{d+2}}$, and the $(\ln K)/T$ term is negligible compared to the other two terms. The rate of convergence then becomes the standard nonparametric rate $N^{-\frac{2}{d+2}}$ of a discrete approximation based on known groups.

Example 3: regression with time-varying unobservables. In order to illustrate the use of Theorem 3, in the appendix we consider the following regression model with additive time-varying unobserved heterogeneity:

$$Y_{it} = X_{it}'\theta_0 + \eta_{i0}(t) + \varepsilon_{it}. \tag{13}$$

This is a model with time-varying paths of heterogeneity $\eta_{i0}(t)$ which may correlate with covariates X_{it} (as in Bonhomme and Manresa, 2015). We derive the rate of convergence of a "double grouped fixed-effects" estimator for this model, whereby we first discretize Y_{it} and all components of X_{it} separately using kmeans clustering, and in a second step regress Y_{it} on X_{it} and all group indicators entering additively, interacted with time indicators.

5.3 Combining panel data and cross-sectional information

A different setting where fixed-effects approaches may not be well-suited is when the longitudinal information on α_i is not sufficient to estimate it in an individual-by-individual fashion. In that case grouped estimators may be useful if some information on α_i can be extracted from a panel sample, although it cannot be directly used to estimate the α_i values.

Specifically, suppose that one has a single direct measure on α_{i0}:

$$W_i = \alpha_{i0} + U_i,$$

and that we have in addition some panel data moments which are informative about α_{i0}, say $h_i = \frac{1}{T}\sum_{t=1}^{T} h(Y_{it}, X_{it})$ with $h_i = \varphi(\alpha_{i0}) + \varepsilon_i$. Note that $(\varphi(\alpha_{i0})', \alpha_{i0}')'$ is a (Lipschitz) function of the scalar α_{i0}, so its underlying dimension is one. Our goal is to estimate α_{i0}. To do so, we combine the two types of information using our two-step approach. In the first step, we apply kmeans clustering to the set of moments $\{h_i\}$, and obtain the partition $\{\widehat{k}_i\}$. In the second step we estimate $\widehat{\alpha}(\widehat{k}_i)$ as the mean of W_i in group $\widehat{k}_i$.

Assumption 7. *(panel and cross-section)*

Observations are i.i.d. across individual units. U_i are independent of α_{i0} and $(\varepsilon_{i1}', ..., \varepsilon_{iT}')'$. U_i has zero mean and finite variance.

Corollary 4. *Let Assumptions 3 and 7 hold. Then as N, T, K tend to infinity:*

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2 = O_p\left(\frac{1}{T}\right)+O_p\left(\frac{K}{N}\right)+O_p\left(B_{\alpha}(K)\right).$$

Applications of this framework include the measurement of latent skills, using a battery of measures, combined with an "anchor" W_i (for example a wage) which expresses the skills in interpretable units; see for example Cunha, Heckman and Schennach (2010). Our illustration on matched employer employee data is another application, where wages of all workers in a firm provide information about firm heterogeneity, but information from job movers (typically a small share of firm employment) is needed to identify the firm's contribution to wages.[13]

6 A dynamic model of location choice

In this section we describe a structural dynamic model of location choice. Discrete estimation improves tractability for three reasons: unobserved state variables have a small number of points of support, the total number of parameters is relatively small, and wage and utility parameters can be estimated sequentially. We present simulations which assess the quality of discrete approximation for two-step, iterative, and bias-reduced grouped fixed-effects estimators.

6.1 Model and estimation method

We consider a model of location choices over J possible alternatives. There is a continuum of agents i who differ in their permanent type $\alpha_i \in \mathbb{R}^J$ which defines their wage in each location. Log-wages in location j (net of age effects and other demographics) are given by: $\ln W_{it}(j) = \alpha_i(j) + \varepsilon_{it}(j)$, where $\varepsilon_{it}(j)$ are assumed to be i.i.d over time, agents, and locations,

[13]In this case, the $1/T$ term in Corollary 4 decreases as the number of workers per firm grows, while the K/N term decreases as the number of job movers increases. In addition, this second term increases with the number of groups K.

distributed as normal $(0, \sigma^2)$. The flow utility of being in location j at time t is given by: $U_{it}(j) = \rho W_{it}(j) + \xi_{it}(j)$, where $\xi_{it}(j)$ are unobserved shocks i.i.d across agents, time and locations, and distributed as type-I extreme value. When moving between two locations, the agent faces a cost c. We denote as $c(j, j') = c\mathbf{1}\{j' \neq j\}$ the cost of moving from j to j'.

Agent i has uncertainty about her own type α_i. While we assume she knows the distribution from which α_i are drawn, she only observes $\alpha_i(j)$ in the locations j she has visited, and she forms expectations about the value she might get in locations she has not visited yet. At time t, let $\mathcal{J}_{it}$ denote the set of locations that agent i has visited. Let $\alpha_i(\mathcal{J}_{it})$ denote the set of corresponding realized location-specific types. The information set of the agent is:

$$S_{it} = \left(j_{it}, \mathcal{J}_{it}, \alpha_i(\mathcal{J}_{it})\right).$$

Note that we assume that wage shocks $\varepsilon_{it}(j)$ do not affect the decision to move to another location. This assumption is useful for tractability, though not essential to our approach.

In this version we consider an infinite horizon environment, where agents discount time at a common rate β. At time t, let $V_t(j, S_{i,t-1})$ denote the expected value function associated with choosing location j given state $S_{i,t-1}$ and behaving optimally in the future. In addition, let us denote the integrated value function as:

$$\overline{V}_t(S_{i,t-1}) = \mathbb{E}\left[\max_{j\in\{1,\dots,J\}} V_t(j, S_{i,t-1}) + \xi_{it}(j) \,\middle|\, S_{i,t-1}\right].$$

By Bellman's principle the alternative-specific value functions are:

$$V_t(j, S_{i,t-1}) = \mathbb{E}\left[U_{it}(j) - c(j_{i,t-1}, j) + \beta\overline{V}_t(S_{it}) \,\middle|\, j_{it} = j, S_{i,t-1}\right],$$

where $S_{it} = (j, \mathcal{J}_{i,t-1} \cup \{j\}, \alpha_i(\mathcal{J}_{i,t-1} \cup \{j\}))$.

From the functional forms we obtain (as in Rust, 1994):

$$\overline{V}_t(S_{i,t-1}) = \ln\left(\sum_{j=1}^{J} \exp V_t(j, S_{i,t-1})\right) + \gamma, \tag{14}$$

where $\gamma \approx .57$ is Euler's constant. Moreover:

$$V_t(j, S_{i,t-1}) = \mathbb{E}\left[\rho\exp\left(\alpha_i(j) + \frac{\sigma^2}{2}\right) - c(j_{i,t-1}, j) + \beta\overline{V}_t\left(j, \mathcal{J}_{i,t-1} \cup \{j\}, \alpha_i(\mathcal{J}_{i,t-1} \cup \{j\})\right) \,\middle|\, S_{i,t-1}\right], \tag{15}$$

where the expectation is taken with respect to the distribution of $\alpha_i(j)$ given $\alpha_i(\mathcal{J}_{i,t-1})$.

Lastly, in period t the conditional choice probabilities are:

$$\Pr\left(j_{it} = j \mid S_{i,t-1}\right) = \frac{\exp V_t(j, S_{i,t-1})}{\sum_{j'=1}^{J} \exp V_t(j', S_{i,t-1})}. \tag{16}$$

Computation. Computation of the solution proceeds in a recursive manner. In the case where all locations have been visited, $\mathcal{J}_{it} = \{1, ..., J\}$ so $S_{it} = (j_{it}, \{1, ..., J\}, \{\alpha_i(1), ..., \alpha_i(J)\})$. Denote the corresponding integrated value function given most recent location j as $\overline{V}^J(i, j)$. From (14) and (15) we have:

$$\overline{V}^J(i,j) = \ln\left(\sum_{j'=1}^{J} \exp\left[\rho \exp\left(\alpha_i(j') + \frac{\sigma^2}{2}\right) - c(j,j') + \beta\overline{V}^J(i,j')\right]\right) + \gamma, \quad j = 1, ..., J.$$

We solve this fixed-point system by successive iterations.

Consider now a case where the agent has visited s states in set $\mathcal{J} \subsetneqq \{1, ..., J\}$, and is currently at location j. Let $\overline{V}^s(i, j, \mathcal{J})$ denote her integrated value function. The latter solves:

$$\begin{aligned}\overline{V}^s(i,j,\mathcal{J}) \quad = \quad & \ln\left(\sum_{j'\notin\mathcal{J}} \exp\left[\mathbb{E}_{\mathcal{J}}\left(\rho \exp\left(\alpha_i(j') + \frac{\sigma^2}{2}\right) - c(j,j') + \beta\overline{V}^{s+1}(i,j',\mathcal{J}\cup\{j'\})\right)\right]\right. \\ & \left. + \sum_{j'\in\mathcal{J}} \exp\left[\rho \exp\left(\alpha_i(j') + \frac{\sigma^2}{2}\right) - c(j,j') + \beta\overline{V}^{s}(i,j',\mathcal{J})\right]\right) + \gamma,\end{aligned}$$

where $\mathbb{E}_{\mathcal{J}}$ is taken with respect to the distribution of $\alpha_i(j')$ given $\alpha_i(\mathcal{J})$. In practice we discretize the values of each $\alpha_i(j)$ on a 50-point grid. In the computation of the fixed points we set a 10^{-11} numerical tolerance.

Estimation. Given an i.i.d sample of wages and locations $(W_{i1}, ..., W_{iT}, j_{i1}, ..., j_{iT})$ we first estimate the realized $\alpha_i(j_{it})$ as follows, using grouped fixed-effects:

$$(\{\widehat{k}_i\}, \widehat{\alpha}) = \underset{(\{k_i\},\alpha)}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(\ln W_{it} - \alpha(k_i, j_{it})\right)^2. \tag{17}$$

In the second step, we maximize the log-likelihood of choices; that is:

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{j=1}^{J} \mathbf{1}\{j_{it} = j\} \ln \Pr\left(j_{it} = j \,|\, j_{i,t-1}, \mathcal{J}_{i,t-1}, \widehat{\alpha}(\widehat{k}_i, \mathcal{J}_{i,t-1}), \theta\right), \tag{18}$$

where θ contains utility and cost parameters (ρ and c), and $\mathcal{J}_{i,t-1}$ denotes the set of locations visited by i until period $t-1$.

We use a steepest ascent algorithm to maximize the objective in (18), analogous to the nested fixed point method of Rust (1994). Alternatively, in this model the CCP method of Hotz and Miller (1988) or the iterative method of Aguirregabiria and Mira (2002) could be used, leading to estimators that are asymptotically equivalent to the solution of (18), although they differ in finite samples. A computational alternative to maximize the objective in (18) is the MPEC method of Su and Judd (2012).

The choice probabilities entering the likelihood are given by an estimated counterpart to (16), where the estimated value functions $\widehat{V}_t\left(j', j_{i,t-1}, \mathcal{J}_{i,t-1}, \widehat{\alpha}(\widehat{k}_i, \mathcal{J}_{i,t-1}), \theta\right)$ solve the system (14)-(15). We estimate the conditional expectation in (15) as a conditional mean given $\alpha_i(\mathcal{J}_{i,t-1})$, based on all job movers from $\mathcal{J}_{i,t-1}$ to $j_{it} = j$. Nonparametric or semi-parametric methods could be used for this purpose. We have used both a Nadaraya Watson kernel estimator and a polynomial series estimator. We use an exponential regression estimator in the illustration below.

Note that agents in the model face uncertainty about future values of realized types $\alpha_i(j)$. The problem is only discretized for estimation purposes. This approach contrasts with a model with discrete types in the population, where after observing her type in one location the agent would in many cases be able to infer her type in all other possible locations. In the present setup, uncertainty about future types diminishes over time as the agent visits more locations, but it does not disappear.

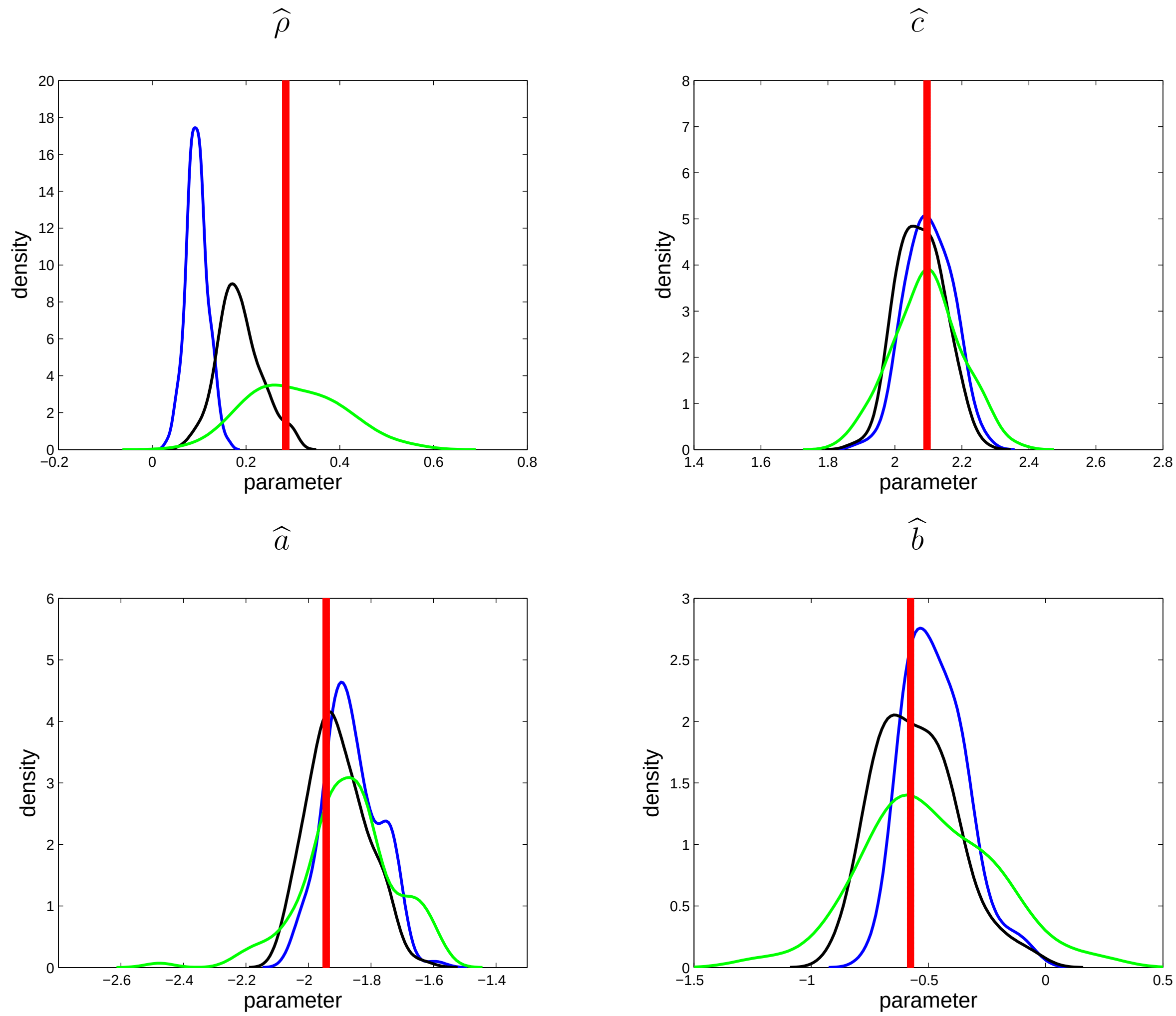
Lastly, given parameter estimates $\widehat{\alpha}(k, j)$, $\widehat{\sigma}^2$, and $\widehat{\theta}$, one can update the estimated partition of individuals using the full model's structure, as in (8) and (9). Details are given in the appendix.

6.2 Simulation exercise

We base our calibration on the NLSY79. We select males who were at least 22 years old in 1979. We keep observations until 1994. Log-wages are regressed on dummies of years of education, race dummies, and a full set of age dummies. We then compute log-wage residuals $\ln W_{it}$. In this version we aggregate regions into $J = 2$ large regions: North-East and South (region A), and North-Central and West (B). There are $N = 1889$ workers, who are observed on average for 12.3 years with a maximum of $T = 16$ years. The probability of moving between the two regions is low in the data: 1.5% per year, only 10.5% of workers moving at all during the observation period. Mean log-wage residuals are .09 higher in region A compared to B.

For estimation we set $K = 10$, and use an iterative estimator starting from a first step based on location-specific mean log-wages. Following Kennan and Walker (2011) we let mobility costs be heterogeneous, individuals being "stayer types" with a probability that we let depend on the initial $\alpha_i(j_{i1})$ though a logistic specification. Hence, while the main parameters of interest are ρ and c, the model also features the intercept and slope coefficients in the probability of being a stayer type (a and b). The estimates we obtain are $\widehat{\rho} = .28$, $\widehat{c} = 2.10$, $\widehat{a} = -1.94$, and $\widehat{b} = -.58$.

Figure 1: Parameter estimates across simulations



Notes: Blue is two-step, black is iterative, green is bias-corrected iterative estimator. The red vertical line indicates the true parameter value. $N = 1889$, $T \leq 16$, $K = 10$. *Unobserved heterogeneity is continuously distributed in the DGP.* 500 *replications.*

In the data generating process (DGP) the probability of being a stayer type is high and depends negatively on the initial location-specific $\alpha_i(j_{i1})$. Costs are also high for non-stayer types. The effect of wages on utility is positive, although we will see below that it is quantitatively small. The model reproduces well the probability of moving, both unconditionally and conditional on past wages; in particular, it reproduces the negative relationship between past wages and mobility. The model also reproduces means and variances of log-wages by location. However, it does not fit well average wages posterior to mobility, as it tend to counterfactually predict mean wage increases upon job move while the data do not show such a pattern.

We next simulate the model based on these parameter values, together with an i.i.d. normal specification of shocks to log wages and the α's. Note that the α's are not discrete in the DGP, although we use a discrete approach in estimation. In Figure 1 we report the results of 500 Monte Carlo simulations. We show three types of density estimates: corresponding to the two-step estimates (blue curve), the iterated estimator (black), and the bias-reduced version of the iterative estimator (in green) based on the split-panel jackknife of Dhaene and Jochmans (2015). The results are shown for the four parameters.

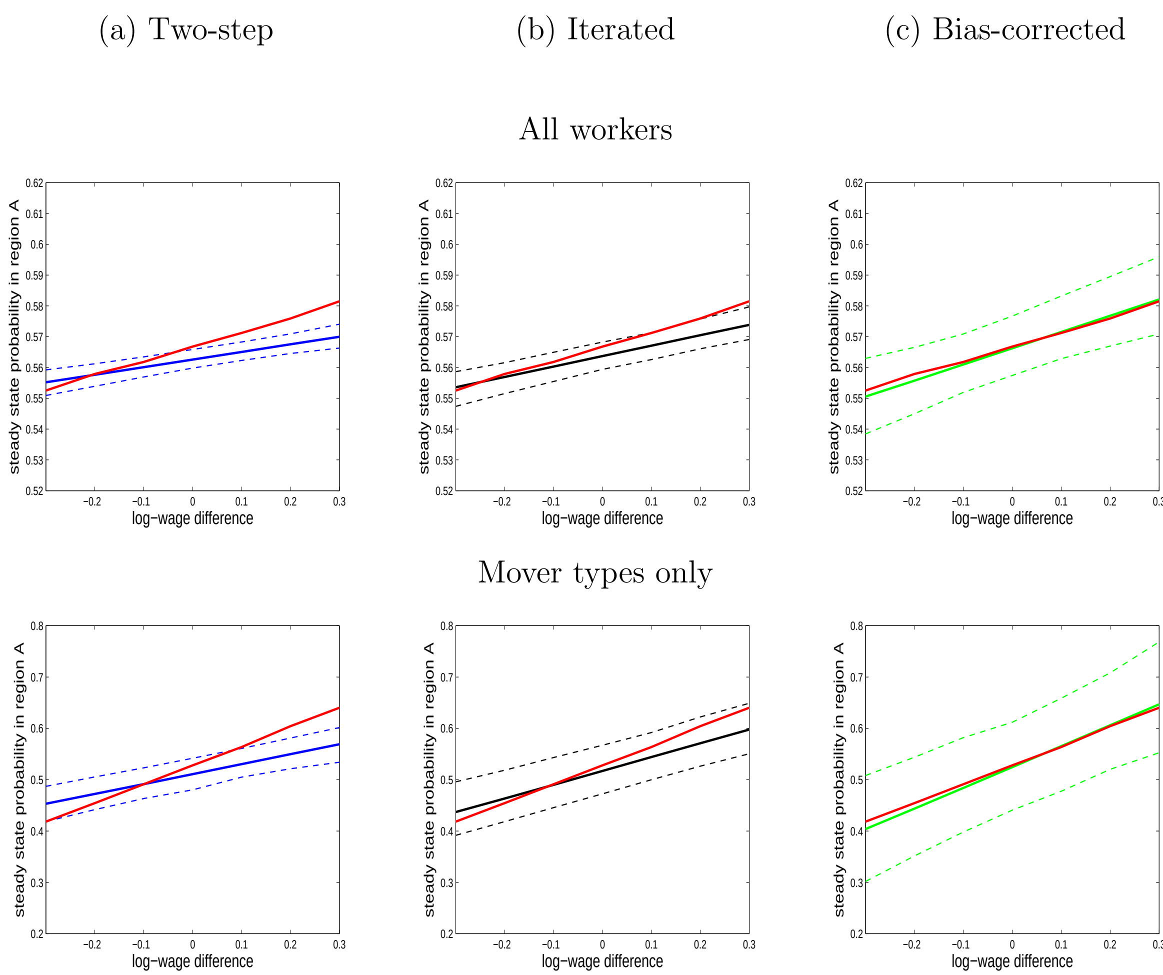
The results on Figure 1 show moderate biases for all three estimators except for the wage effect ρ. In the latter case, using both iteration and bias reduction improves the performance of the estimator substantially. At the same time, on these data bias reduction tends to be associated with a variance increase.

Long-run effects of wages. As an additional exercise, we compute the steady-state probability of working in region A when varying the log-wage differential between A and B. This provides a simple example of a counterfactual experiment that the model can be used for. In Figure 2 we show the log-wage differential on the x-axis, and the probability of working in A on the y-axis. The red curves on the graphs show the estimates from NLSY, while the blue, black and green curves who means and 95% pointwise bands across simulations for the two-step, iterative, and bias-corrected estimators, respectively.[14]

The results on Figure 2 show that the model predicts small effects of wages on mobility on average. When increasing the wage in A relative to B by 30% the probability of working in A increases by less than 2 percentage points, from 56.7% to 58.2%. When focusing on workers who are not of a "stayer type" (bottom panel), whose mobility may be affected by the change in wages, we see a more substantial effect, as increasing the wage in region A by 30% increases the long-run probability of working in A from 52.8% to 64.0%. In both cases the two-step estimator and the iterated estimator are biased downward, the latter being closer to the true value. The bias-corrected estimator is close to unbiased. However it is less precisely estimated, reflecting the estimates of the wage effect ρ on utility (see Figure 1).

[14]In the counterfactual we keep the probability of being a "stayer type" constant. Hence we abstract from the fact that wage increases could affect the distribution of mobility costs, in addition to the effect on utility that we focus on.

Figure 2: Long-run effects of wages



Notes: Blue is two-step, black is iterative, green is bias-corrected iterative estimator. The red curve indicates the true parameter value. Difference in log-wage between the two regions (x-axis), and steady-state probability of working in region A (y-axis). Solid curves are means, and dashed curves are 97.5% and 2.5% percentiles, across simulations. 500 *replications.*

7 Firm and worker heterogeneity

In the second illustration we consider a model of log-wages as an additive combination of worker and firm determinants:

$$Y_{it} = \eta_i + \psi_{j(i,t)} + \varepsilon_{it}, \tag{19}$$

where worker i works in firm $j(i,t)$ at time t, and η_i and ψ_j denote unobserved attributes of worker i and firm j, respectively. Equation (19) corresponds to the model of Abowd, Kramarz and Margolis (1999) for matched employer-employee data, where we abstract from observed covariates for simplicity.

We focus on a simple specification where ε_{it} is independent of $j(i,1),...,j(i,T)$ and of η's and ψ's, and i.i.d. across i,t. Bonhomme, Lamadon and Manresa (2015, BLM hereafter) consider more general models that allow for worker-firm interaction effects and dynamics. In this illustration we purposely focus on a simplified setup to study the biases of grouped fixed-effects estimators in various DGPs.

We estimate the model on two periods. Following BLM we adopt a correlated random-effects approach to model worker heterogeneity within and across firms. The parameters of interest include the firm fixed-effects ψ_j, the means and variances of worker effects in each firm $\mu_j = \mathbb{E}\left(\eta_i \,|\, j(i,1)=j\right)$ and $\sigma_j^2 = \mathrm{Var}\left(\eta_i \,|\, j(i,1)=j\right)$, and the variances of idiosyncratic errors $s_j^2 = \mathrm{Var}\left(\varepsilon_{i1} \,|\, j(i,1)=j\right)$. From these parameters we can decompose the variance of log-wages into a firm component, a worker component, a component refecting the sorting of workers into heterogeneous firms, and an idiosyncratic match component:

$$\mathrm{Var}\left(y_{i1}\right) = \mathrm{Var}\left(\eta_i\right) + \mathrm{Var}\left(\psi_{j(i,1)}\right) + 2\,\mathrm{Cov}\left(\eta_i, \psi_{j(i,1)}\right) + \mathrm{Var}\left(\varepsilon_{i1}\right). \tag{20}$$

We will be estimating the components in this decomposition. In addition we will report estimates of the correlation between worker and firm effects, $\mathrm{Corr}\left(\eta_i, \psi_{j(i,1)}\right)$, which is commonly interpreted as a measure of sorting.

We start by estimating model (19) on Swedish register data. Sample selection follows BLM: we select male workers full-year employed in 2002 and 2004. Job movers are defined as workers whose firm IDs change between the two periods. We focus on firms that are present throughout the period. There are about 20,000 job movers in the sample. We compute residuals from a regression of log-wages on interactions of education and age indicators.

Fixed-effects estimation of model (19) is potentially subject to incidental parameter bias. This is due to the fact that identification of firm effects ψ_j comes from job movements. When the number of job movers into and out of firm j is low, ψ_j may thus be poorly estimated. This situation is more likely to arise in short panels. To address this problem, BLM proposed a two-step estimation approach. In the first step, firms are partitioned into K groups, $\widehat{k}_j \in \{1, ..., K\}$, based on moments of firm-specific log-wage distributions. Here we use log-wage means and standard deviations in the first period, and use the *kmeans* estimator with $K = 10$. In the second step, the model's parameters are estimated at the group level; that is, we estimate $\widehat{\psi}(\widehat{k}_j)$, $\mu(\widehat{k}_j)$, $\sigma^2(\widehat{k}_j)$, and $s^2(\widehat{k}_j)$. The second step relies on simple mean and covariance restrictions, as we describe in the appendix.

We next simulate two periods of wages from the estimated model. We consider two DGPs: a first one with continuous heterogeneity, where we simulate ψ_j from a normal distribution, calibrated to match the mean and variance of $\widehat{\psi}(\widehat{k}_j)$,[15] and a second DGP with discrete heterogeneity where all firm population parameters are constant within groups $\widehat{k}_j$ (with $K = 10$). We draw 120,000 workers in the cross-section, including 20,000 job movers. We run simulations for different firm sizes, from 5 workers per firm to 100 workers per firm.

In the DGP, firm heterogeneity is four-dimensional, but its underlying dimension is one. Indeed, the vector of firm-specific parameters is:

$$\alpha_j = \left(\psi_j \, , \, \mu_j \, , \, \sigma_j^2, s_j^2 \right) = \left(\psi_j, \, \mathbb{E}\left(\eta_i | \psi_{j(i,1)} = \psi_j \right) \, , \, \mathrm{Var}\left(\eta_i | \psi_{j(i,1)} = \psi_j \right) \, , \, \mathrm{Var}\left(\varepsilon_{i1} | \psi_{j(i,1)} = \psi_j \right) \right).$$

In other words, all firm-specific parameters are (nonlinear) functions of the scalar firm effects ψ_j. This assumption is consistent with theoretical models of worker-firm sorting where firms are characterized by their scalar productivity level, as in Shimer and Smith (2000) for example.[16]

In Table 1 we report the mean and variance of parameters in the variance decomposition (20) across 100 simulations of the continuous DGP. In Figure 3 we report means and 95% pointwise confidence bands across simulations. We see that two-step estimators are biased for small firms, particularly the variance of firm effects and the correlation between worker and firm effects. For example, the bias is 68% for the former and 49% for the latter when firm size equals 5, and biases are still 33% and 14% when firm size equals 20. In the two bottom panels of Table 1 (and the right panel of Figure 3) we show the results using the bias-correction method of Dhaene and Jochmans (2015).[17] The bias-corrected estimator tends to perform better: biases are 53% and 20% for the variance of firm effects and the correlation between worker and firm effects when firm size equals 5, and they reduce to 13% and 2% when firm size equals 20. Moreover,

[15] Likewise, we draw the ψ_{j_1} and ψ_{j_2} of job movers from a bivariate normal distribution.

[16] On the other hand, the presence of a non-pecuniary firm-specific amenity A_j could lead to a higher underlying dimension, as in this case the mean and variance of η_i would generally be functions of (ψ_j, A_j).

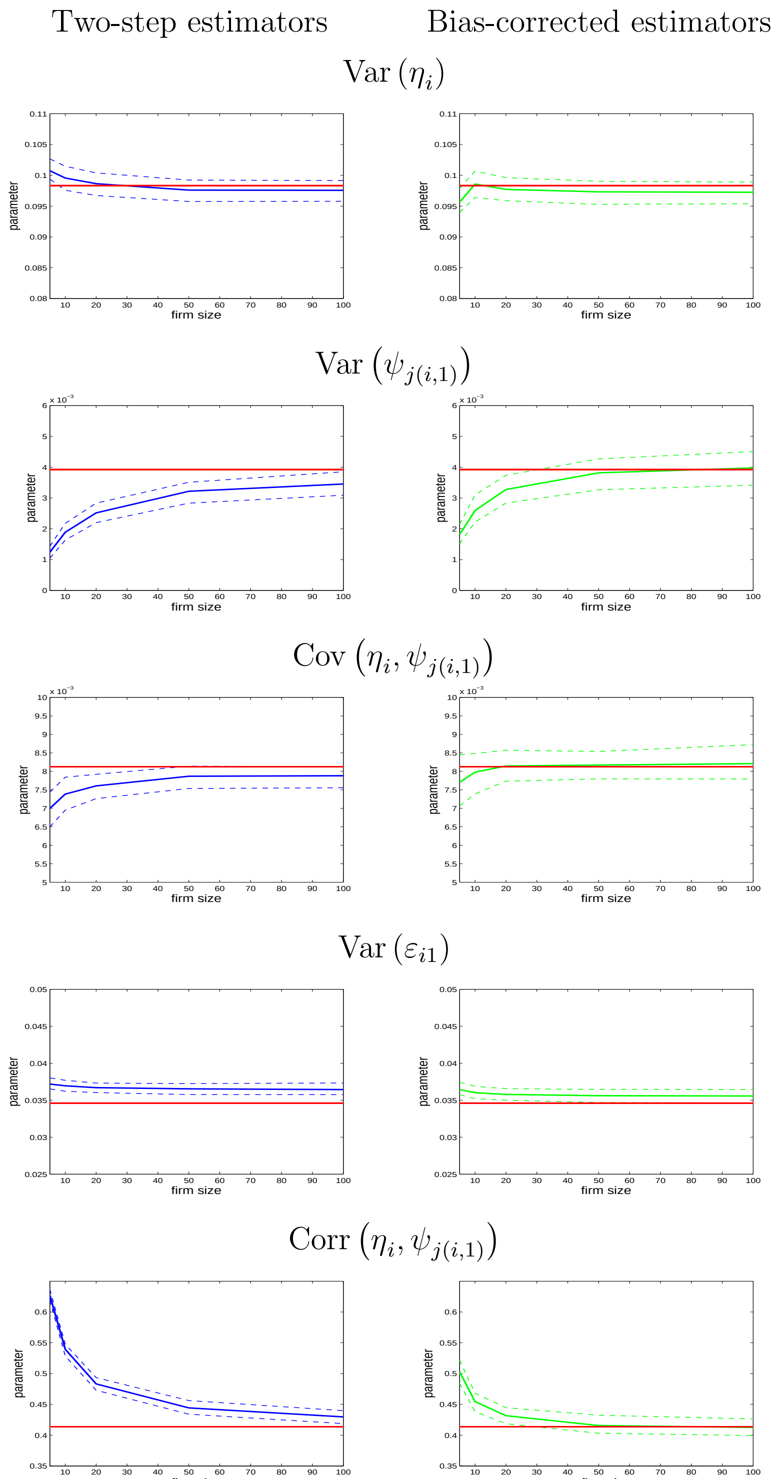
[17] To implement the method we select two halves within each firm at random. We take $7 \approx K/\sqrt{2}$ groups in each half-sample. Note that one could alternatively average across random permutations.

Table 1: Firm and worker effects, continuous data generating process

	$\text{Var}(\eta_i)$	$\text{Var}(\psi_{j(i,1)})$	$\text{Cov}(\eta_i, \psi_{j(i,1)})$	$\text{Var}(\varepsilon_{i1})$	$\text{Corr}(\eta_i, \psi_{j(i,1)})$
	True values				
	.0983	.0039	.0081	.0346	.4137
	Two-step estimator, means				
Firm size = 5	.1007	.0012	.0070	.0372	.6270
Firm size = 10	.0996	.0019	.00739	.0369	.5395
Firm size = 20	.0986	.0025	.0076	.0367	.4833
Firm size = 50	.0976	.0032	.0079	.0365	.4443
Firm size = 100	.0976	.0035	.0079	.0364	.4297
	Two-step estimator, standard deviations				
Firm size = 5	.00086	.00011	.00027	.00039	.00504
Firm size = 10	.00104	.00015	.00022	.00041	.00510
Firm size = 20	.00100	.00016	.00017	.00036	.00498
Firm size = 50	.00086	.00019	.00015	.00041	.00565
Firm size = 100	.00090	.00021	.00015	.00044	.00583
	Bias-corrected estimator, means				
Firm size = 5	.0957	.0018	.0077	.0364	.5027
Firm size = 10	.0986	.0026	.0080	.0360	.4545
Firm size = 20	.0977	.0033	.0081	.0358	.4316
Firm size = 50	.0973	.0038	.0082	.0356	.4156
Firm size = 100	.0972	.0040	.0082	.0356	.4127
	Bias-corrected estimator, standard deviations				
Firm size = 5	.00108	.00018	.00036	.00044	.00968
Firm size = 10	.00115	.00023	.00028	.00042	.00769
Firm size = 20	.00106	.00024	.00023	.00040	.00683
Firm size = 50	.00095	.00026	.00020	.00046	.00753
Firm size = 100	.00096	.00029	.00021	.00049	.00737

Notes: Mean and standard deviation of parameters across simulations. $K = 10$. 100 *simulations.*

Figure 3: Firm and worker heterogeneity, continuous DGP



Notes: Means (solid) and 95% confidence bands (dashed). Blue is two-step, green is bias-corrected estimator. The red curve indicates the true parameter value. 100 *replications.*

bias correction is not associated with large increases in dispersion. Hence this approach seems to provide a clear improvement in this setting.

Discrete population heterogeneity. In Table D1 in the appendix we show the same results as in Table 1, now for a discrete DGP generated with $K_0 = 10$ groups. Figure D4 in the appendix plots means and 95% confidence bands. The results turn out to be very similar to those obtained in the continuous DGP. In particular, for small firm size we see substantial biases for some parameters. This suggests that for this DGP the "oracle" asymptotic theory based on the premise that group misclassification is irrelevant in the limit does not provide reliable guidance for finite sample inference. In fact, in this discrete DGP we computed group misclassification frequencies across simulations, and found 76% misclassification when firm size equals 5, 61% when size is 20, and still 40% when firm size is 100. Interestingly, in this discrete setting too, we find that bias correction improves the performance of the estimator.

8 Conclusion

In this paper we analyze properties of discrete heterogeneity methods while being agnostic about the form of individual unobserved heterogeneity. We show that grouped fixed-effects estimators are subject to an approximation bias, which comes from the population not being discrete, and an incidental parameter bias, which stems from groups being estimated with noise. We show in two illustrations that fixed-effects bias correction methods can be successfully adapted to improve the performance of discrete estimators.

Grouped fixed-effects methods may be particularly useful when unobservables are multi-dimensional, provided their underlying dimension is low. A case in point is a model with time-varying unobservables following a low-dimensional nonlinear factor structure. A question which we do not address is how to infer the underlying dimension from the data. We do not address either the questions of the choice of the number of groups and the characterization of asymptotic distributions in high-dimensional settings. These questions are important for future work.

Finally, two-step and iterative grouped fixed-effects methods could be useful beyond our two empirical illustrations. Other settings include models with multi-sided heterogeneity, nonlinear factor models, semi-parametric panel data models such as quantile regression with individual effects, and network models, for example. We also see grouped fixed-effects methods as potentially useful in structural models such as dynamic discrete choice. In models with multiple locations, a curse of dimensionality arises when workers keep track of their full history of locations (as in Kennan and Walker, 2011). Extending clustering methods to address this dimensionality challenge would be interesting.

References

[1] Abowd, J., F. Kramarz, and D. Margolis (1999): "High Wage Workers and High Wage Firms", *Econometrica*, 67(2), 251–333.

[2] Aguirregabiria, V., and P. Mira (2002): "Swapping the Nested Fixed-Point Algorithm: A Class of Estimators for Discrete Markov Decision Models," *Econometrica*, 70(4), 1519–1543.

[3] Aguirregabiria, V., and P. Mira (2010): "Dynamic discrete choice structural models: A survey," *Journal of Econometrics*, 156, 38–67.

[4] Arcidiacono, P., and J. B. Jones (2003): 'Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm", *Econometrica*, 71(3), 933–946.

[5] Arcidiacono, P., and R. Miller (2011): 'Conditional Choice Probability Estimation of Dynamic Discrete Choice Models With Unobserved Heterogeneity", *Econometrica*, 79(6), 1823–1867.

[6] Arellano, M. and B. Honoré (2001): "Panel Data Models: Some Recent Developments", in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, vol. 5, North Holland, Amsterdam.

[7] Arellano, M., and J. Hahn (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments,". In: R. Blundell, W. Newey, and T. Persson (eds.): *Advances in Economics and Econometrics, Ninth World Congress*, Cambridge University Press.

[8] Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171.

[9] Bai, J. (2009), "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77, 1229–1279.

[10] Bai, J., and T. Ando (2015):"Panel Data Models with Grouped Factor Structure Under Unknown Group Membership," *Journal of Applied Econometrics.*

[11] Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

[12] Bester, A., and C. Hansen (2016): "Grouped Effects Estimators in Fixed Effects Models", *Journal of Econometrics*, 190(1), 197–208.

[13] Bickel, P.J., and A. Chen (2009): "A Nonparametric View of Network Models and Newman-Girvan and Other Modularities," *Proc. Natl. Acad. Sci. USA*, 106, 21068–21073.

[14] Bonhomme, S., and E. Manresa (2015): "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 83(3), 1147–1184.

[15] Bonhomme, S., T. Lamadon, and E. Manresa (2015): "A Distributional Framework for Matched Employer-Employee Data," unpublished manuscript.

[16] Bryant, P. and Williamson, J. A. (1978): "Asymptotic Behaviour of Classification Maximum Likelihood Estimates," *Biometrika*, 65, 273–281.

[17] Chamberlain, G., and M. Rotschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51(5), 1281–1304.

[18] Choi, D.S., P.J. Wolfe, and E.M. Airoldi (2012): "Stochastic Blockmodels with a Growing Number of Classes," *Biometrika*, 99, 273–284.

[19] Cunha, F., J. Heckman, and S. Schennach (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formation", *Econometrica*, 78(3), 883–931.

[20] Dhaene, G. and K. Jochmans (2015): "Split Panel Jackknife Estimation," to appear in *Review of Economic Studies*.

[21] Gao, C., Y. Lu, and H. H. Zhou (2015): "Rate-Optimal Graphon Estimation," *Annals of Statistics*, 43(6), 2624–2652.

[22] Gersho, A., and R.M. Gray (1992): *Vector Quantization and Signal Compression*. Kluwer Academic Press.

[23] Graf, S., and H. Luschgy (2000): *Foundations of Quantization for Probability Distributions.* Springer Verlag, Berlin, Heidelberg.

[24] Graf, S., and H. Luschgy (2002): "Rates of Convergence for the Empirical Quantization Error", *Annals of Probability.*

[25] Gray, R. M., and D. L. Neuhoff (1998): "Quantization", *IEEE Trans. Inform. Theory,* (Special Commemorative Issue), 44(6), 2325–2383.

[26] Hahn, J. and W.K. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models", *Econometrica*, 72, 1295–1319.

[27] Hahn, J., and H. Moon (2010): "Panel Data Models with Finite Number of Multiple Equilibria," *Econometric Theory*, 26(3), 863–881.

[28] Hastie, T., and W. Stuetzle (1989): "Principal Curves," *Journal of the American Statistical Association*, 84, 502–516.

[29] Heckman, J.J., and B. Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52(2), 271–320.

[30] Hotz, J., and R. Miller (1988): "Conditional Choice Probabilities and the Estimation of Dynamic Models", *Review of Economic Studies*, 60(3), 497–529.

[31] Hsu, D., S.M. Kakade, and T. Zhang (2012): "A Tail Inequality for Quadratic Forms of Subgaussian Random Vectors," *Electron. Commun. Probab.*, 17, 52, 1–6.

[32] Keane, M., and K. Wolpin (1997): "The Career Decisions of Young Men," *Journal of Political Economy*, 105(3), 473–522.

[33] Kennan, J., and J. Walker (2011): "The Effect of Expected Income on Individual Migration Decisions", *Econometrica*, 79(1), 211–251.

[34] Levina, E., and P. J. Bickel (2004): "Maximum Likelihood Estimation of Intrinsic Dimension," *Advances in neural information processing systems*, 777–784.

[35] Linder, T. (2002): *Learning-Theoretic Methods in Vector Quantization*, Principles of Nonparametric Learning. L. Gy¨orfi, editor, CISM Lecture Notes, Wien, New York.

[36] Lin, C. C., and S. Ng (2012): "Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown", *Journal of Econometric Methods*, 1(1), 42–55.

[37] Pollard, D. (1981): "Strong Consistency of K-means Clustering," *Annals of Statistics*, 9, 135– 140.

[38] Pollard, D. (1982a): "A Central Limit Theorem for K-Means Clustering," *Annals of Probability*, 10, 919–926.

[39] Pollard, D. (1982b): "Quantization and the Method of K-Means," *IEEE Transactions on Information Theory*, vol. IT-28 (2), 199–205.

[40] Raginsky, M., and Lazebnik, S. (2005): "Estimation of Intrinsic Dimensionality Using High-Rate Vector Quantization," *In Advances in neural information processing systems*, 1105–1112.

[41] Rust, J. (1994): "Structural Estimation of Markov Decision Processes," *Handbook of econometrics*, 4(4), 3081–3143.

[42] Saggio, R. (2012): "Discrete Unobserved Heterogeneity in Discrete Choice Panel Data Models," CEMFI Master Thesis.

[43] Shimer, R., and L. Smith (2000): "Assortative Matching and Search," *Econometrica*, 68(2), 343–369.

[44] Steinley, D. (2006): "K-means Clustering: A Half-Century Synthesis," *Br. J. Math. Stat. Psychol.*, 59, 1–34.

[45] Su, C. and K. Judd (2012): "Constrained Optimization Approaches to Estimation of Structural Models", *Econometrica*, 80(5), 2213–2230.

[46] Su, L., Z. Shi, and P. C. B. Phillips (2015): "Identifying Latent Structures in Panel Data," unpublished manuscript.

[47] Vershynin, R. (2010): *Introduction to the Non-Asymptotic Analysis of Random Matrices*, in Y. C. Eldar and G. Kutyniok, ed., Compressed Sensing: Theory and Applications. Cambridge University Press.

[48] Vogt and O. Linton (2015): "Classification of Nonparametric Regression Functions in Heterogeneous Panels," unpublished manuscript.

[49] Wolfe, P. J., and S. C. Ohlede (2014): "Nonparametric Graphon Estimation", Arkiv.

APPENDIX

A Proofs

A.1 Proof of Theorem 1

Let us define:[18]

$$(\widetilde{\varphi}, \{\widetilde{k}_i\}) = \underset{(\varphi,\{k_i\})}{\operatorname{argmin}} \sum_{i=1}^{N} \|\varphi(\alpha_{i0}) - \varphi(k_i)\|^2 . \tag{A1}$$

By definition of $(\widehat{\varphi}, \{\widehat{k}_i\})$, we have (with probability one):

$$\sum_{i=1}^{N} \left\| h_i - \widehat{\varphi}(\widehat{k}_i) \right\|^2 \leq \sum_{i=1}^{N} \left\| h_i - \widetilde{\varphi}(\widetilde{k}_i) \right\|^2 .$$

Letting $\varepsilon_i = \sum_{t=1}^{T} \varepsilon_{it}/T$, with $\varepsilon_{it} = h(Y_{it}, X_{it}) - \varphi(\alpha_{i0})$, we thus have:

$$\frac{1}{N}\sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widehat{\varphi}(\widehat{k}_i) \right\|^2 \quad \leq \quad \underbrace{\frac{1}{N}\sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widetilde{\varphi}(\widetilde{k}_i) \right\|^2}_{=B_\varphi(K)} + \frac{2}{N}\sum_{i=1}^{N} \varepsilon_i' \left(\widehat{\varphi}(\widehat{k}_i) - \widetilde{\varphi}(\widetilde{k}_i) \right) .$$

Moreover, by the Cauchy-Schwarz and triangular inequalities:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i' \left(\widehat{\varphi}(\widehat{k}_i) - \widetilde{\varphi}(\widetilde{k}_i) \right) \quad &\leq \quad \left(\frac{1}{N}\sum_{i=1}^{N} \|\varepsilon_i\|^2 \right)^{\frac{1}{2}} \cdot \left(\frac{1}{N}\sum_{i=1}^{N} \left\| \widehat{\varphi}(\widehat{k}_i) - \widetilde{\varphi}(\widetilde{k}_i) \right\|^2 \right)^{\frac{1}{2}} \\
&\leq \quad \left(\frac{1}{N}\sum_{i=1}^{N} \|\varepsilon_i\|^2 \right)^{\frac{1}{2}} \cdot \left(\left(\frac{1}{N}\sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widehat{\varphi}(\widehat{k}_i) \right\|^2 \right)^{\frac{1}{2}} + B_\varphi(K)^{\frac{1}{2}} \right) .
\end{aligned}$$

Letting $E^2 = \frac{1}{N}\sum_{i=1}^{N} \|\varepsilon_i\|^2$, it thus follows that:

$$\frac{1}{N}\sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widehat{\varphi}(\widehat{k}_i) \right\|^2 - B_\varphi(K) \leq 2E \cdot \left(\left(\frac{1}{N}\sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widehat{\varphi}(\widehat{k}_i) \right\|^2 \right)^{\frac{1}{2}} + B_\varphi(K)^{\frac{1}{2}} \right),$$

so, dividing both sides by $\left(\frac{1}{N}\sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widehat{\varphi}(\widehat{k}_i) \right\|^2 \right)^{\frac{1}{2}} + B_\varphi(K)^{\frac{1}{2}}$:

$$\left(\frac{1}{N}\sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widehat{\varphi}(\widehat{k}_i) \right\|^2 \right)^{\frac{1}{2}} \leq B_\varphi(K)^{\frac{1}{2}} + 2E.$$

Hence, as $E^2 = O_p(1/T)$ and $B_\varphi(K) = O_p(B_\alpha(K))$ by Assumption 1, we have:

$$\frac{1}{N}\sum_{i=1}^{N} \left\| \varphi(\alpha_{i0}) - \widehat{\varphi}(\widehat{k}_i) \right\|^2 = O_p\left(\frac{1}{T} \right) + O_p(B_\alpha(K)).$$

[18] The literature on vector quantization provides general results on existence of optimal empirical quantizers; that is, solutions to (A1). See for example Chapter 1 in Graf and Luschgy (2000).

A.2 Proof of Corollary 1

Let $\{k_i\}$ be the intersection of two partitions of $\{1,...,N\}$: a first partition with K_1 groups, and a second partition with (the integer part of) K/K_1 groups. We have:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left\|h_i-\widehat{\varphi}(\widehat{k}_i)\right\|^2 &= \frac{1}{N}\sum_{i=1}^{N}\left\|\varphi(\alpha_{i0})+\varepsilon_i-\widehat{\varphi}(\widehat{k}_i)\right\|^2 \\ &\leq B_{\varphi}(K_1)+\frac{1}{T}\cdot B_{\sqrt{T}\varepsilon}\left(\frac{K}{K_1}\right) \\ &= O_p(K_1^{-\frac{2}{d}})+\frac{1}{T}\cdot O_p\left(\left(\frac{K}{K_1}\right)^{-\frac{2}{r}}\right).\end{aligned}$$

Letting $K_1=(TK^{\frac{2}{r}})^{\frac{dr}{2(d+r)}}$ (or its integer part), and noting that K/K_1 tends to infinity due to $TK^{-\frac{2}{d}}$ tending to zero, yields the desired result.

A.3 Proof of Theorem 2

Define the fixed-effects estimator of α_i, for given θ, as $\widehat{\alpha}_i(\theta)=\operatorname{argmax}_{\alpha_i}\ \ell_i(\alpha_i,\theta)$. Define the "target" log-likelihood estimand (e.g., Arellano and Hahn, 2007) as $\overline{\theta}=\operatorname{argmax}_{\theta}\sum_{i=1}^{N}\ell_i(\overline{\alpha}_i(\theta),\theta)$, where $\overline{\alpha}_i(\theta)=\operatorname{argmax}_{\alpha_i}\ \operatorname{plim}_{T\to\infty}\ell_i(\alpha_i,\theta)$. Note that $\overline{\alpha}_i(\theta_0)=\alpha_{i0}$. Moreover, under Assumption 2:

$$\overline{\theta}=\theta_0+H^{-1}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\ell_i(\alpha_{i0},\theta_0)}{\partial\theta}+o_p\left(\frac{1}{\sqrt{NT}}\right).$$

Let also, for all θ and $k\in\{1,...,K\}$:

$$\widehat{\alpha}(k,\theta)=\underset{\alpha}{\operatorname{argmax}}\ \sum_{i=1}^{N}\mathbf{1}\{\widehat{k}_i=k\}\ell_i\left(\alpha,\theta\right). \tag{A2}$$

We start by noting that there exists a Lipschitz-continuous function $\overline{\alpha}$ such that $\overline{\alpha}_i(\theta)=\overline{\alpha}\left(\alpha_{i0},\theta\right)$ for all i and θ. This follows from the implicit functions theorem, as $\operatorname{plim}_{T\to\infty}\frac{\partial\ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i}=0$ and, by part (ii) in Assumption 3, for all (α,θ) $\operatorname{plim}_{T\to\infty}\frac{\partial\ell_i(\alpha,\theta)}{\partial\alpha_i}$ is only a function of α_{i0}.

Moreover, for all θ:

$$\sum_{i=1}^{N}\ell_i\left(a(\widehat{k}_i,\theta),\theta\right)\leq\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)\leq\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i(\theta),\theta\right),$$

where $a(k,\theta)=\overline{\alpha}\left(\psi(\widehat{\varphi}(k)),\theta\right)$.

Now, by a second order expansion and using Assumption 2:

$$\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(a(\widehat{k}_i,\theta),\theta\right)-\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i(\theta),\theta\right)\right|=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|a(\widehat{k}_i,\theta)-\widehat{\alpha}_i(\theta)\right\|^2\right).$$

Moreover, by Assumptions 2 and 3:

$$\begin{aligned}
\frac{1}{2N}\sum_{i=1}^{N}\left\|a(\widehat{k}_i,\theta)-\widehat{\alpha}_i(\theta)\right\|^2 &\leq \frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\alpha}\left(\psi(\widehat{\varphi}(\widehat{k}_i)),\theta\right)-\overline{\alpha}\left(\alpha_{i0},\theta\right)\right\|^2+\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\alpha}\left(\alpha_{i0},\theta\right)-\widehat{\alpha}_i(\theta)\right\|^2 \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\alpha}\left(\psi(\widehat{\varphi}(\widehat{k}_i)),\theta\right)-\overline{\alpha}\left(\alpha_{i0},\theta\right)\right\|^2+\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\alpha}_i\left(\theta\right)-\widehat{\alpha}_i(\theta)\right\|^2 \\
&= O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\psi(\widehat{\varphi}(\widehat{k}_i))-\psi\left(\varphi(\alpha_{i0})\right)\right\|^2\right)+O_p\left(\frac{1}{T}\right) \\
&= O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\varphi}(\widehat{k}_i)-\varphi(\alpha_{i0})\right\|^2\right)+O_p\left(\frac{1}{T}\right).
\end{aligned}$$

Hence, using Assumption 1 we obtain, using Theorem 1:

$$\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(a(\widehat{k}_i,\theta),\theta\right)-\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i(\theta),\theta\right)\right|=O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right),$$

from which it follows that:

$$\left|\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)-\frac{1}{N}\sum_{i=1}^{N}\ell_i\left(\widehat{\alpha}_i(\theta),\theta\right)\right|=O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right).$$

As a result, by a second-order expansion, and using concavity in Assumption 3, we obtain:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i,\theta)-\widehat{\alpha}_i(\theta)\right\|^2=O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right). \tag{A3}$$

We also obtain that $\widehat{\theta}\xrightarrow{p}\theta_0$ as N,T,K tend to infinity. More precisely, this argument shows that $\|\widehat{\theta}-\theta_0\|^2=O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right)$. Hence, in particular, (6) follows.

In the rest of the proof we establish (5). The proof proceeds by bounding the score difference between the two-step grouped fixed-effects estimator and the target estimator; that is:

$$\Delta S(\theta)=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\left(\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)-\ell_i\left(\overline{\alpha}_i(\theta),\theta\right)\right).$$

When $\Delta S(\theta_0)=O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right)+o_p\left(\frac{1}{\sqrt{NT}}\right)$ has been established, (5) will follow from standard arguments.

From (A2) and Assumption 2 it follows that:

$$\Delta S(\theta)=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)}{\partial\theta}-\frac{\partial\ell_i\left(\overline{\alpha}_i(\theta),\theta\right)}{\partial\theta}-\frac{\partial\overline{\alpha}_i(\theta)'}{\partial\theta}\frac{\partial\ell_i\left(\overline{\alpha}_i(\theta),\theta\right)}{\partial\alpha_i},$$

where we note that, by (A2), $\sum_{i=1}^{N}\frac{\partial\widehat{\alpha}(\widehat{k}_i,\theta)'}{\partial\theta}\frac{\partial\ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)}{\partial\alpha_i}=0$.

Moreover, a second-order expansion gives:

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ell_i\left(\widehat{\alpha}(\widehat{k}_i,\theta),\theta\right)}{\partial\theta}-\frac{\partial \ell_i\left(\overline{\alpha}_i(\theta),\theta\right)}{\partial\theta} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \ell_i\left(\overline{\alpha}_i(\theta),\theta\right)}{\partial\theta\partial\alpha_i'}\left(\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right) +O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right),$$

where we have used Assumption 2, and that by (A3):

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\alpha}(\widehat{k}_i,\theta)-\overline{\alpha}_i(\theta)\right\|^2=O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right). \tag{A4}$$

Now, by Assumption 3 there exists a function $\nabla_{\theta\alpha}\ell$ such that, at true parameter values:

$$\text{plim}_{T\rightarrow\infty}\frac{\partial^2 \ell_i\left(\alpha_{i0},\theta_0\right)}{\partial\theta\partial\alpha_i'}=\nabla_{\theta\alpha}\ell\left(\alpha_{i0}\right).$$

Hence, using (A4), Assumption 2, and the Cauchy Schwarz inequality:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \ell_i\left(\alpha_{i0},\theta_0\right)}{\partial\theta\partial\alpha_i'}\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\overline{\alpha}_i(\theta_0)\right) &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \ell_i\left(\alpha_{i0},\theta_0\right)}{\partial\theta\partial\alpha_i'}\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\alpha_{i0}\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta\alpha}\ell\left(\alpha_{i0}\right)\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\alpha_{i0}\right)\\
&\quad +O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right)\\
&= \frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta\alpha}\ell\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)\right)\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\widehat{\alpha}_i(\theta_0)\right)\\
&\quad +\frac{1}{N}\sum_{i=1}^{N}\nabla_{\theta\alpha}\ell\left(\alpha_{i0}\right)\left(\widehat{\alpha}_i(\theta_0)-\alpha_{i0}\right)\\
&\quad +O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right).\end{aligned}$$

Note that, from $\frac{\partial \ell_i(\widehat{\alpha}_i(\theta_0),\theta_0)}{\partial\alpha_i}=0$, and from the fact that $\text{plim}_{T\rightarrow\infty}\frac{\partial \ell_i(\overline{\alpha}_i(\theta),\theta)}{\partial\alpha_i}=0$ for all θ, we have using Assumption 2:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\nabla_{\theta\alpha}\ell\left(\alpha_{i0}\right)\left(\widehat{\alpha}_i(\theta_0)-\alpha_{i0}\right)-\frac{\partial\overline{\alpha}_i(\theta_0)'}{\partial\theta}\frac{\partial \ell_i\left(\alpha_{i0},\theta_0\right)}{\partial\alpha_i}\right) = O_p\left(\frac{1}{T}\right).$$

Moreover, for all $k\in\{1,...,K\}$ we have, by (A2):

$$\sum_{j=1}^{N}\mathbf{1}\{\widehat{k}_j=k\}\frac{\partial \ell_j\left(\widehat{\alpha}(k,\theta_0),\theta_0\right)}{\partial\alpha_j}=0.$$

Hence, expanding around $\widehat{\alpha}_j(\theta_0)$:

$$\widehat{\alpha}(k,\theta_0)=\left(\sum_{j=1}^N \mathbf{1}\{\widehat{k}_j=k\}\frac{\partial^2\ell_j\left(b_j(k),\theta_0\right)}{\partial\alpha_j\partial\alpha_j'}\right)^{-1}\sum_{j=1}^N \mathbf{1}\{\widehat{k}_j=k\}\frac{\partial^2\ell_j\left(b_j(k),\theta_0\right)}{\partial\alpha_j\partial\alpha_j'}\widehat{\alpha}_j(\theta_0), \tag{A5}$$

where $b_j(k)$ lies between $\widehat{\alpha}_j(\theta_0)$ and $\widehat{\alpha}(k,\theta_0)$.

We also have:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \nabla_{\theta\alpha}\ell\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)\right)\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\widehat{\alpha}_i(\theta_0)\right) &= \frac{1}{N}\sum_{i=1}^N \nabla_{\theta\alpha}\ell\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)\right)\left(\frac{\partial^2\ell_i\left(b_i(\widehat{k}_i),\theta_0\right)}{\partial\alpha_i\partial\alpha_i'}\right)^{-1} \\ &\quad\times\left(\frac{\partial^2\ell_i\left(b_i(\widehat{k}_i),\theta_0\right)}{\partial\alpha_i\partial\alpha_i'}\right)\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\widehat{\alpha}_i(\theta_0)\right).\end{aligned} \tag{A6}$$

Now, by Assumption 3 there exists a function $\nabla_{\alpha\alpha}\ell$ such that:

$$\begin{aligned}\text{plim}_{T\to\infty}\frac{\partial^2\ell_i\left(\alpha_{i0},\theta_0\right)}{\partial\alpha_i\partial\alpha_i'} &= \nabla_{\alpha\alpha}\ell(\alpha_{i0}) \\ &= \nabla_{\alpha\alpha}\ell(\widehat{\alpha}(\widehat{k}_i,\theta_0))+O_p\left(\left\|\widehat{\alpha}(\widehat{k}_i,\theta_0)-\alpha_{i0}\right\|\right).\end{aligned}$$

It thus follows from Assumption 2 that:

$$\frac{\partial^2\ell_i\left(b_i(\widehat{k}_i),\theta_0\right)}{\partial\alpha_i\partial\alpha_i'}=\nabla_{\alpha\alpha}\ell(\widehat{\alpha}(\widehat{k}_i,\theta_0))+O_p\left(\left\|\widehat{\alpha}(\widehat{k}_i,\theta_0)-\alpha_{i0}\right\|+\|\widehat{\alpha}_i(\theta_0)-\alpha_{i0}\|\right)+O_p\left(\frac{1}{\sqrt{T}}\right).$$

Hence, using (A3) and (A6), we have:

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^N \nabla_{\theta\alpha}\ell\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)\right)\left(\frac{\partial^2\ell_i\left(b_i(\widehat{k}_i),\theta_0\right)}{\partial\alpha_i\partial\alpha_i'}\right)^{-1}\left(\frac{\partial^2\ell_i\left(b_i(\widehat{k}_i),\theta_0\right)}{\partial\alpha_i\partial\alpha_i'}\right)\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\widehat{\alpha}_i(\theta_0)\right) \\ &\quad=\underbrace{\frac{1}{N}\sum_{i=1}^N \nabla_{\theta\alpha}\ell\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)\right)\left(\nabla_{\alpha\alpha}\ell(\widehat{\alpha}(\widehat{k}_i,\theta_0))\right)^{-1}\left(\frac{\partial^2\ell_i\left(b_i(\widehat{k}_i),\theta_0\right)}{\partial\alpha_i\partial\alpha_i'}\right)\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\widehat{\alpha}_i(\theta_0)\right)}_{=0 \text{ by (A5)}} \\ &\quad\quad+O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right).\end{aligned}$$

So:

$$\frac{1}{N}\sum_{i=1}^N \nabla_{\theta\alpha}\ell\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)\right)\left(\widehat{\alpha}(\widehat{k}_i,\theta_0)-\widehat{\alpha}_i(\theta_0)\right) = O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right).$$

This establishes (5) and ends the proof of Theorem 2.

A.4 Proof of Corollary 2

We have, by Theorem 2:

$$\begin{aligned}
\widehat{M} - M_0 &= \frac{1}{N}\sum_{i=1}^{N} m_i\left(\widehat{\alpha}(\widehat{k}_i), \widehat{\theta}\right) - \frac{1}{N}\sum_{i=1}^{N} m_i\left(\alpha_{i0}, \theta_0\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} \frac{\partial m_i\left(\alpha_{i0}, \theta_0\right)}{\partial \alpha_i'}\left(\widehat{\alpha}(\widehat{k}_i, \widehat{\theta}) - \alpha_{i0}\right) + \frac{1}{N}\sum_{i=1}^{N} \frac{\partial m_i\left(\alpha_{i0}, \theta_0\right)}{\partial \theta'}\left(\widehat{\theta} - \theta_0\right) \\
&\quad + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) + o_p\left(\frac{1}{\sqrt{NT}}\right).
\end{aligned}$$

Hence, from the proof of Theorem 2, and using Assumption 4:

$$\begin{aligned}
\widehat{M} - M_0 &= \frac{1}{N}\sum_{i=1}^{N} \frac{\partial m_i\left(\alpha_{i0}, \theta_0\right)}{\partial \alpha_i'}\left(\widehat{\alpha}_i(\overline{\theta}) - \alpha_{i0}\right) + \frac{1}{N}\sum_{i=1}^{N} \frac{\partial m_i\left(\alpha_{i0}, \theta_0\right)}{\partial \theta'}\left(\overline{\theta} - \theta_0\right) \\
&\quad + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) + o_p\left(\frac{1}{\sqrt{NT}}\right).
\end{aligned}$$

The result then comes from substituting $\overline{\theta}-\theta_0$ by its influence function, and expanding $\frac{\partial \ell_i(\widehat{\alpha}_i(\theta),\theta)}{\partial \alpha_i} = 0$ around $\overline{\alpha}_i(\theta)$.

A.5 Proof of Corollary 3

We start by noting that, by definition of $\{\widehat{k}_i^{(2)}\}$:

$$\sum_{i=1}^{N} \ell_i\left(\widehat{\alpha}(\widehat{k}_i), \widehat{\theta}\right) \leq \sum_{i=1}^{N} \ell_i\left(\widehat{\alpha}(\widehat{k}_i^{(2)}), \widehat{\theta}\right).$$

Hence, a second-order expansion around $\widehat{\alpha}(\widehat{k}_i)$ gives:

$$\sum_{i=1}^{N} \frac{\partial \ell_i\left(\widehat{\alpha}(\widehat{k}_i), \widehat{\theta}\right)}{\partial \alpha_i'}\left(\widehat{\alpha}(\widehat{k}_i^{(2)}) - \widehat{\alpha}(\widehat{k}_i)\right) + \frac{1}{2}\sum_{i=1}^{N}\left(\widehat{\alpha}(\widehat{k}_i^{(2)}) - \widehat{\alpha}(\widehat{k}_i)\right)' \frac{\partial^2 \ell_i\left(a_i, \widehat{\theta}\right)}{\partial \alpha_i \partial \alpha_i'}\left(\widehat{\alpha}(\widehat{k}_i^{(2)}) - \widehat{\alpha}(\widehat{k}_i)\right) \geq 0,$$

where a_i lies between $\widehat{\alpha}(\widehat{k}_i)$ and $\widehat{\alpha}(\widehat{k}_i^{(2)})$.

Hence, using Assumption 3, Theorem 2, and the Cauchy-Schwarz inequality, we obtain:

$$\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\alpha}(\widehat{k}_i^{(2)}) - \widehat{\alpha}(\widehat{k}_i)\|^2 \leq O_p\left(\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial \ell_i\left(\widehat{\alpha}(\widehat{k}_i), \widehat{\theta}\right)}{\partial \alpha_i}\right\|^2\right)^{\frac{1}{2}} \cdot \left(\frac{1}{N}\sum_{i=1}^{N}\|\widehat{\alpha}(\widehat{k}_i^{(2)}) - \widehat{\alpha}(\widehat{k}_i)\|^2\right)^{\frac{1}{2}}\right),$$

from which it follows, using Theorem 2, that:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \|\widehat{\alpha}(\widehat{k}_i^{(2)}) - \widehat{\alpha}(\widehat{k}_i)\|^2 &\leq O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{\partial \ell_i\left(\alpha_{i0}, \theta_0\right)}{\partial \alpha_i}\right\|^2\right) + O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right) \\
&= O_p\left(\frac{1}{T}\right) + O_p\left(B_\alpha(K)\right).
\end{aligned}$$

Hence:

$$\frac{1}{N}\sum_{i=1}^N \|\widehat{\alpha}(\widehat{k}_i^{(2)})-\alpha_{i0}\|^2 \quad = \quad O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right).$$

This establishes that there exists a function of $\{\widehat{k}_i^{(2)}\}$ which approximates the true α_{i0} on average at the desired rate.

The end of the proof is as in the proof of Theorem 2.

A.6 Bias of the one-step estimator in Example 2

The derivations in this subsection are heuristic. Consider model (4), that is: $Y_{it}=W_{it}'\theta_0+\eta_{i0}+\varepsilon_{it}$. Pollard (1981, 1982a) provides conditions under which, for fixed K,T and as N tends to infinity, the one-step grouped fixed-effects estimator is root-N consistent and asymptotically normal for the minimizer θ^* of the following population objective function:[19]

$$Q(\theta)=\underset{N\rightarrow\infty}{\text{plim}}\ \underset{(\eta,\{k_i\})}{\text{argmin}}\ \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\left(Y_{it}-W_{it}'\theta-\eta(k_i)\right)^2.$$

Now:

$$Q(\theta)=\underset{N\rightarrow\infty}{\text{plim}}\ \frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\left(Y_{it}-\overline{Y}_i-\left(W_{it}-\overline{W}_i\right)'\theta\right)^2+Q_B(\theta),$$

where:

$$Q_B(\theta)=\underset{N\rightarrow\infty}{\text{plim}}\ \underset{(\eta,\{k_i\})}{\text{argmin}}\ \frac{1}{N}\sum_{i=1}^N\left(\overline{Y}_i-\overline{W}_i'\theta-\eta(k_i)\right)^2.$$

From Theorem 6.2 in Graf and Luschgy (2000) we have, as K tends to infinity for fixed T, and provided the density f_θ of $\overline{Y}_i-\overline{W}_i'\theta$ is non-singular with respect to the Lebesgue measure:

$$Q_B(\theta)=\frac{1}{12K^2}\left(\int\left[f_\theta(y)\right]^{\frac{1}{3}}dy\right)^3+o\left(\frac{1}{K^2}\right).$$

As an example, consider the case where observations are i.i.d across i, and $\overline{Y}_i-\overline{W}_i'\theta$ is normal with mean $\mu(\theta)$ and variance $\sigma^2(\theta)$. Then calculations show that:

$$\frac{1}{12K^2}\left(\int\left[f_\theta(y)\right]^{\frac{1}{3}}dy\right)^3=\frac{\pi\sqrt{3}}{2K^2}\sigma^2(\theta).$$

Moreover:

$$\frac{\partial\sigma^2(\theta)}{\partial\theta}=2\,\text{Var}(\overline{W}_i)\theta-2\,\text{Cov}\left(\overline{W}_i,\overline{Y}_i\right).$$

[19] Pollard focuses on the standard kmeans estimator, without covariates. See the supplementary appendix in Bonhomme and Manresa (2015) for an analysis with covariates.

This suggests that, up to an $o(K^{-2})$ term, the pseudo-true value θ^* solves:

$$\mathbb{E}\left[-\frac{1}{T}\sum_{t=1}^{T}(W_{it}-\overline{W}_i)\left(Y_{it}-\overline{Y}_i-(W_{it}-\overline{W}_i)'\theta\right)\right]+\frac{\pi\sqrt{3}}{K^2}\left(\mathrm{Var}(\overline{W}_i)\theta-\mathrm{Cov}\left(\overline{W}_i,\overline{Y}_i\right)\right)=0.$$

This gives:

$$\begin{aligned}\theta^* &= \left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(W_{it}-\overline{W}_i)\left(W_{it}-\overline{W}_i\right)'\right]+\frac{\pi\sqrt{3}}{K^2}\mathrm{Var}(\overline{W}_i)\right)^{-1}\\ &\quad\times\left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(W_{it}-\overline{W}_i)\left(Y_{it}-\overline{Y}_i\right)\right]+\frac{\pi\sqrt{3}}{K^2}\mathrm{Cov}\left(\overline{W}_i,\overline{Y}_i\right)\right)+o\left(\frac{1}{K^2}\right).\end{aligned}$$

Hence:

$$\begin{aligned}\theta^*-\theta_0 &= \left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(W_{it}-\overline{W}_i)\left(W_{it}-\overline{W}_i\right)'\right]+\frac{\pi\sqrt{3}}{K^2}\mathrm{Var}(\overline{W}_i)\right)^{-1}\\ &\quad\times\left(\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(W_{it}-\overline{W}_i)\,\varepsilon_{it}\right]+\frac{\pi\sqrt{3}}{K^2}\mathrm{Cov}\left(\overline{W}_i,\eta_i+\overline{\varepsilon}_i\right)\right)+o\left(\frac{1}{K^2}\right).\end{aligned}$$

A.7 Proof of Theorem 3

Following the steps of the proof of Theorem 1, it is easy to show that, as N,T,K,q tend to infinity:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\varphi(\alpha_{i0})-\widehat{\varphi}(\widehat{k}_i)\right\|^2=O_p\left(\frac{1}{N}\sum_{i=1}^{N}\|\varepsilon_i\|^2\right)+O_p\left(B_\varphi(K)\right),$$

where by Assumption 6 $B_\varphi(K)=O_p(B_\alpha(K))$, and by Assumption 5 $\frac{1}{N}\sum_{i=1}^N\|\varepsilon_i\|^2=O_p(\sigma_T^2)$.

In the rest of the proof we establish the other part of the bound.

Let $\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)$ denote the linear projection of ε_i on the indicators $\mathbf{1}\{\widehat{k}_i=k\}$ and $\mathbf{1}\{\widetilde{k}_i=k\}$, all of which are interacted with component indicators. Following similar steps as in the proof of Theorem 1, we obtain:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left\|\varphi(\alpha_{i0})-\widehat{\varphi}(\widehat{k}_i)\right\|^2 &\leq B_\varphi(K)+\frac{2}{N}\sum_{i=1}^{N}\varepsilon_i'\left(\widehat{\varphi}(\widehat{k}_i)-\widetilde{\varphi}(\widetilde{k}_i)\right)\\ &= B_\varphi(K)+\frac{2}{N}\sum_{i=1}^{N}\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)'\left(\widehat{\varphi}(\widehat{k}_i)-\widetilde{\varphi}(\widetilde{k}_i)\right)\\ &\leq B_\varphi(K)+O_p\left(\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)\right\|^2\right)^{\frac{1}{2}}\cdot\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\varphi(\alpha_i)-\widehat{\varphi}(\widehat{k}_i)\right\|^2\right)^{\frac{1}{2}}\right).\end{aligned}$$

Hence, using once again arguments from the proof of Theorem 1, and by Assumption 6:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\varphi(\alpha_{i0})-\widehat{\varphi}(\widehat{k}_i)\right\|^2 = O_p(B_\alpha(K))+O_p\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)\right\|^2\right).$$

Finally, in order to show that:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)\right\|^2 = O_p\left(c_T\cdot\ln K\right)+O_p\left(c_T\cdot\frac{qK}{N}\right), \tag{A7}$$

we apply a version the Hanson-Wright tail inequality for quadratic forms, due to Hsu, Kakade and Zhang (2012, Theorem 2.1), which allows for dependent data.

Lemma A1. *(Hsu et al., 2012) Let Z be a m-dimensional random vector such that $\mathbb{E}\left[\exp(\alpha' Z)\right]\leq \exp(c\left\|\alpha\right\|^2)$, for all $\alpha\in\mathbb{R}^m$. Let A be a positive semi-definite matrix. Then, for all $s>0$:*

$$\Pr\left[Z'AZ>2c\,\mathrm{tr}\,A+4c\sqrt{s\,\mathrm{tr}\,A^2}+s4c\left\|A\right\|\right]\leq\exp(-s).$$

Let, for given partitions $\{k_{i1}\},\{k_{i2}\}$:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(k_{1i},k_{2i})\right\|^2=\frac{\varepsilon' A\varepsilon}{N},$$

where $\varepsilon=(\varepsilon_1',...,\varepsilon_N')'$, A is a $rN\times rN$ projection matrix with $\mathrm{tr}\,A=2rK$, $A^2=A$, and $\|A\|=1$. By Lemma A1 and Assumption 5 we have:

$$\Pr\left[\varepsilon' A\varepsilon>4c_T rK+4c_T\sqrt{2rKs}+4c_T s\right]\leq\exp(-s),$$

so, using that $2\sqrt{ab}\leq a+b$:

$$\Pr\left[\varepsilon' A\varepsilon>8c_T rK+6c_T s\right]\leq\exp(-s),$$

hence:

$$\Pr\left[\frac{\varepsilon' A\varepsilon}{N}>b\right]\leq\exp\left[-\left(\frac{bN}{6c_T}-\frac{4rK}{3}\right)\right].$$

Lastly, by the union bound, given that the set of partitions $\{k_{i1}\}\cap\{k_{i2}\}$ has K^{2N} elements:

$$\begin{aligned}\Pr\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(\widehat{k}_i,\widetilde{k}_i)\right\|^2>b\right] &\leq K^{2N}\max_{\{k_{i1}\},\{k_{i2}\}}\Pr\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\varepsilon}(k_{1i},k_{2i})\right\|^2>b\right]\\ &\leq \exp\left[2N\ln K+\frac{4rK}{3}-\frac{bN}{6c_T}\right].\end{aligned}$$

Using that r/q tends to a positive constant, this shows (A7) and ends the proof.

A.8 Example 3: regression with time-varying unobservables

In the next result we derive the rate of convergence of a "double grouped fixed-effects" estimator in model (13). Let η_0 be a vector containing all $\eta_{i0}(t)$ for $i=1,...,N$ and $t=1,...,T$. Define μ_0 similarly below.

Assumption A1. *(covariates)*

(i) $X_{it} = \mu_{i0}(t) + V_{it}$, *where* $(V_{11}', ..., V_{NT}')'$ *satisfies part (ii) in Assumption 5 conditional on* η_0 *and* μ_0.

(ii) $\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T V_{it}V_{it}' \xrightarrow{p} \Sigma > 0$ *as* N, T *tend to infinity.*

(iii) As $N, T \to \infty$: $\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T V_{it}\varepsilon_{it} = O_p\left(\frac{1}{\sqrt{NT}}\right)$, $\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T V_{it}\eta_{i0}(t) = O_p\left(\frac{1}{\sqrt{NT}}\right)$, *and* $\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T V_{it}\mu_{i0}(t)' = O_p\left(\frac{1}{\sqrt{NT}}\right)$.

Corollary A1. *Let part (ii) in Assumption 5 hold conditional on* η_0 *and* μ_0, *for* $(\varepsilon_{11}, ..., \varepsilon_{NT})'$. *Let Assumption A1 hold. Let* $B(K) = \max\{B_\eta(K), B_{\mu_1}(K), ..., B_{\mu_R}(K)\}$, *where* R *denote the number of components of the vector* X_{it}.

Then, as N, T, K *tend to infinity such that* $(\ln K)/T$ *and* K/N *tend to zero, the double grouped fixed-effects estimator* $\widehat{\theta}$ *of* θ_0 *satisfies:*

$$\widehat{\theta} = \theta_0 + O_p\left(\frac{\ln K}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(\frac{B(K)}{T}\right) + O_p\left(\frac{1}{\sqrt{NT}}\right).$$

Proof. Let $\widetilde{X}_{it}$ denote the linear projection of X_{it} on all group indicators, entering additively and interacted with time indicators, where the group indicators are obtained by discretizing Y_{it} and the components of of X_{it}. Let $\widetilde{Y}_{it}$ denote the projection of Y_{it} on the same set of indicators. The double grouped fixed-effects estimator of θ_0 is:

$$\begin{aligned}
\widehat{\theta} &= \left[\sum_{i=1}^N \sum_{t=1}^T \left(X_{it} - \widetilde{X}_{it}\right)\left(X_{it} - \widetilde{X}_{it}\right)'\right]^{-1} \sum_{i=1}^N \sum_{t=1}^T \left(X_{it} - \widetilde{X}_{it}\right)\left(Y_{it} - \widetilde{Y}_{it}\right) \\
&= \theta_0 + \left[\sum_{i=1}^N \sum_{t=1}^T \left(X_{it} - \widetilde{X}_{it}\right)\left(X_{it} - \widetilde{X}_{it}\right)'\right]^{-1} \sum_{i=1}^N \sum_{t=1}^T \left(X_{it} - \widetilde{X}_{it}\right)\left(\eta_{i0}(t) - \widetilde{\eta}_{i0}(t) + \varepsilon_{it} - \widetilde{\varepsilon}_{it}\right).
\end{aligned}$$

We start with a lemma.

Lemma A2. *All the following terms are* $O_p\left(\frac{\ln K}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(\frac{B(K)}{T}\right)$ *as* N, T, K *tend to infinity:*

$$A = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(V_{it} - \widetilde{V}_{it}\right)\left(\mu_{i0}(t) - \widetilde{\mu}_{i0}(t)\right)', \quad B = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(\mu_{i0}(t) - \widetilde{\mu}_{i0}(t)\right)\left(\mu_{i0}(t) - \widetilde{\mu}_{i0}(t)\right)',$$

$$C = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(\eta_{i0}(t) - \widetilde{\eta}_{i0}(t)\right)^2, \quad D = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(V_{it} - \widetilde{V}_{it}\right)\left(\eta_{i0}(t) - \widetilde{\eta}_{i0}(t)\right),$$

$$E = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \widetilde{V}_{it}\widetilde{V}_{it}', \quad F = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \widetilde{\varepsilon}_{it}^2, \quad G = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(\varepsilon_{it} - \widetilde{\varepsilon}_{it}\right)\left(\mu_{i0}(t) - \widetilde{\mu}_{i0}(t)\right)'.$$

Proof. B and C are bounded by the approximation bias $B(K)/T$. E and F are bounded as in the proof of Theorem 3, using part (ii) in Assumption 5. Finally, to bound A, D and G we use a similar argument. Let us focus on A for conciseness. We have:

$$\begin{aligned} A &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}\left(\mu_{i0}(t)-\widetilde{\mu}_{i0}(t)\right)' = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}\mu_{i0}(t)' - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}\widetilde{\mu}_{i0}(t)' \\ &= O_p\left(\frac{1}{\sqrt{NT}}\right) - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}\widetilde{\mu}_{i0}(t)'. \end{aligned}$$

Recall that $\widetilde{V}_{it}$ and $\widetilde{\mu}_{i0}(t)$ are constructed using estimated partitions. Following the logic of the proof of Theorem 3 we can consider analogous linear projections on the infeasible partitions based on the true $\eta_{i0}(t)$ and $\mu_{i0}(t)$. In the proof of Theorem 3 these were denoted as $\widetilde{k}_i$, and were infeasible counterparts to the estimated groups $\widehat{k}_i$. Let us denote as $\mu^*_{i0}(t)$ the linear projections of $\mu_{i0}(t)$ on all these infeasible group indicators, entering additively and interacted with time indicators. Using similar arguments in the proof of Theorem 3, we have:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}\left(\widetilde{\mu}_{i0}(t)-\mu^*_{i0}(t)\right)' = O_p\left(\frac{\ln K}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(\frac{B(K)}{T}\right).$$

So, using part (i) in Assumption A1:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}\widetilde{\mu}_{i0}(t)' = O_p\left(\frac{\ln K}{T}\right) + O_p\left(\frac{K}{N}\right) + O_p\left(\frac{B(K)}{T}\right).$$

This completes the proof of Lemma A2.

∎

Throughout the rest of the proof of Corollary A1 we use the notation from Lemma A2. We first have:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\widetilde{X}_{it}\right)\left(X_{it}-\widetilde{X}_{it}\right)' - \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}V_{it}' = A + A' + B - E,$$

so $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\widetilde{X}_{it}\right)\left(X_{it}-\widetilde{X}_{it}\right)' \xrightarrow{p} \Sigma > 0$.

Next, using the Cauchy-Schwarz and triangular inequalities:

$$\begin{aligned} \left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\widetilde{X}_{it}\right)\left(\eta_{i0}(t)-\widetilde{\eta}_{i0}(t)\right)\right\| &\leq \left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\mu_{i0}(t)-\widetilde{\mu}_{i0}(t)\right)\left(\eta_{i0}(t)-\widetilde{\eta}_{i0}(t)\right)\right. \\ &\qquad \left. + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(V_{it}-\widetilde{V}_{it}\right)\left(\eta_{i0}(t)-\widetilde{\eta}_{i0}(t)\right)\right\| \\ &\leq \sqrt{\|B\|\cdot C} + \|D\|, \end{aligned}$$

and:

$$\begin{aligned} \left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}-\widetilde{X}_{it}\right)\left(\varepsilon_{it}-\widetilde{\varepsilon}_{it}\right)\right\| &\leq \|G\| + \left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\widetilde{V}_{it}\widetilde{\varepsilon}_{it}\right\| + \left\|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}V_{it}\varepsilon_{it}\right\| \\ &\leq \|G\| + \sqrt{\|E\|\cdot F} + O_p\left(\frac{1}{\sqrt{NT}}\right). \end{aligned}$$

This ends the proof of Corollary A1. ∎

A.9 Proof of Corollary 4

Similarly as in the proofs of Theorem 1 and Theorem 2 we obtain that:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\psi(\widehat{\alpha}(\widehat{k}_i))-\alpha_{i0}\right\|^2=O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right),$$

where ψ is given in Assumption 3.

As a result, denoting as $\overline{\alpha}(k)$ the mean of α_i in group $\widehat{k}_i=k$, we have:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{\alpha}(\widehat{k}_i)-\alpha_{i0}\right\|^2=O_p\left(\frac{1}{T}\right)+O_p\left(B_\alpha(K)\right).$$

Finally, note that $\widehat{\alpha}(\widehat{k}_i)=\overline{\alpha}(\widehat{k}_i)+\overline{U}(\widehat{k}_i)$, where $\overline{U}(k)$ denotes the mean of U_i in group $\widehat{k}_i=k$. By Assumption 7, we have:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\overline{U}(\widehat{k}_i)\right\|^2\leq \mathrm{Var}(U_i)\cdot\frac{K}{N}.$$

B Complements to the applications

B.1 Dynamic model of location choice

Iterated grouped fixed-effects estimation. To proceed, we first estimate the idiosyncratic variance of log-wages σ^2 as:

$$\widehat{\sigma}^2=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\ln W_{it}-\widehat{\alpha}(\widehat{k}_i,j_{it})\right)^2. \quad \text{(B8)}$$

Then, individual groups are assigned as:

$$\widehat{k}_i^{(2)} = \underset{k\in\{1,...,K\}}{\mathrm{argmax}}\sum_{t=1}^{T}\sum_{j=1}^{J}\mathbf{1}\{j_{it}=j\}\Bigg(\ln\Pr\left(j_{it}=j\,|\,j_{i,t-1},\mathcal{J}_{i,t-1},\widehat{\alpha}(k,\mathcal{J}_{i,t-1}),\widehat{\theta}\right) + \ln\phi(\ln W_{it};\widehat{\alpha}(k,j),\widehat{\sigma}^2)\Bigg),$$

where ϕ denotes the normal density. Note that information on both wages and choices is used to reclassify individuals.

Given group assignments, parameters can be updated as:

$$\widehat{\alpha}^{(2)}(k,j)=\frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{\widehat{k}_i^{(2)}=k\}\mathbf{1}\{j_{it}=j\}\ln W_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{\widehat{k}_i^{(2)}=k\}\mathbf{1}\{j_{it}=j\}},$$

with an update for σ^2 analogous to (B8), and:

$$\widehat{\theta}^{(2)} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{j=1}^{J} \mathbf{1}\{j_{it}=j\} \ln \Pr\left(j_{it}=j \,|\, j_{i,t-1}, \mathcal{J}_{i,t-1}, \widehat{\alpha}^{(2)}(\widehat{k}_i^{(2)}, \mathcal{J}_{i,t-1}), \theta\right).$$

This procedure may be iterated. Note that in the update step we do not maximize the full likelihood as a function of parameters α, σ^2, θ. Rather, we use a sequential likelihood estimator by which we first estimate wage parameters α and σ^2, and then estimate utility and cost parameters θ. We use this approach for computational reasons; see Rust (1994) and Arcidiacono and Jones (2003) for related approaches.

B.2 Firm and worker heterogeneity

Following BLM we exploit the following restrictions, where we denote as $m_i = 1$ the mobility indicator. For job movers, using the fact that mobility does not depend on ε's, and that ε_{i1} is independent of ε_{i2}, we have:

$$\mathbb{E}\left(Y_{i2} - Y_{i1} \,|\, m_i = 1, \psi_{j(i,1)}, \psi_{j(i,2)}\right) = \psi_{j(i,2)} - \psi_{j(i,1)}, \tag{B9}$$

$$\begin{aligned} \operatorname{Var}\left(Y_{i2} - Y_{i1} \,|\, m_i = 1, \psi_{j(i,1)}, \psi_{j(i,2)}\right) &= \operatorname{Var}\left(\varepsilon_{i2} \,|\, \psi_{j(i,2)}\right) + \operatorname{Var}\left(\varepsilon_{i1} \,|\, \psi_{j(i,1)}\right) \\ &= s^2_{j(i,2)} + s^2_{j(i,1)}. \end{aligned} \tag{B10}$$

Then, in the first cross-section we have:

$$\begin{aligned} \mathbb{E}\left(Y_{i1} \,|\, \psi_{j(i,1)}\right) &= \psi_{j(i,1)} + \mathbb{E}\left(\eta_i \,|\, \psi_{j(i,1)}\right) \\ &= \psi_{j(i,1)} + \mu_{j(i,1)}, \end{aligned} \tag{B11}$$

$$\begin{aligned} \operatorname{Var}\left(Y_{i1} \,|\, \psi_{j(i,1)}\right) &= \operatorname{Var}\left(\eta_i \,|\, \psi_{j(i,1)}\right) + \operatorname{Var}\left(\varepsilon_{i1} \,|\, \psi_{j(i,1)}\right) \\ &= \sigma^2_{j(i,1)} + s^2_{j(i,1)}. \end{aligned} \tag{B12}$$

In estimation, we first compute a firm partition $\{\widehat{k}_j\}$ into K groups based on firm-specific log-wage means and standard deviations. In the second step, we use the following algorithm:

1. Compute $\widehat{\psi}(\widehat{k}_j)$ based on mean restrictions for job movers; see (B9).

2. Compute $\widehat{s}^2(\widehat{k}_j)$ based on variance restrictions for job movers; see (B10). In practice we impose non-negativity of the parameters using a quadratic programming routine.

3. From $\widehat{\psi}(\widehat{k}_j)$, compute $\widehat{\mu}(\widehat{k}_j)$ based on mean restrictions in the first cross-section; see (B11).

4. From $\widehat{s}^2(\widehat{k}_j)$, compute $\widehat{\sigma}^2(\widehat{k}_j)$ based on variance restrictions in the first cross-section; see (B12). We also impose non-negativity.

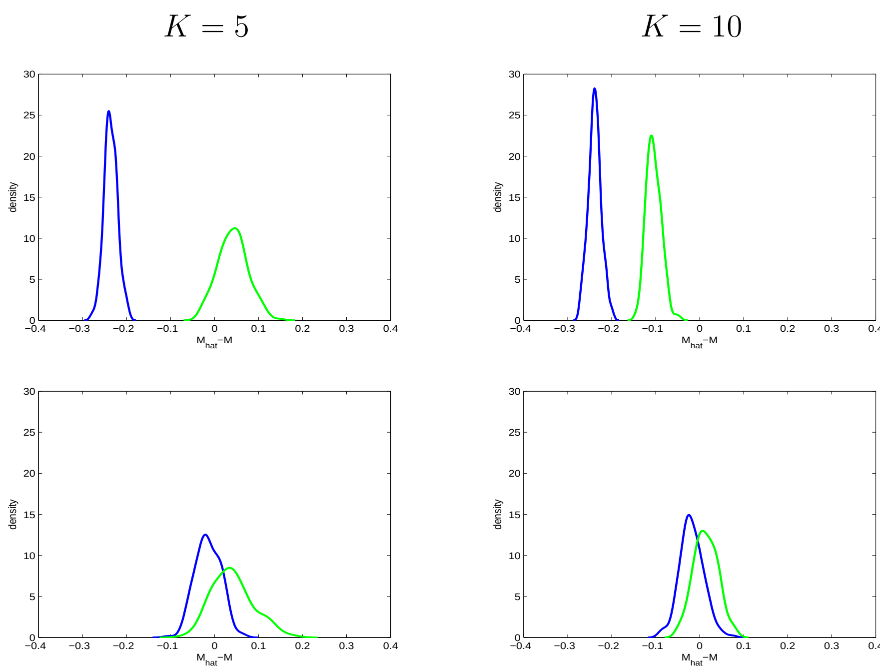
Lastly, given parameter estimates, we compute the variances and covariances in (20) by aggregation across types.

C Additional simulation exercises

C.1 A heterogeneous AR(1) model

Consider the model $Y_{it} = \rho_i Y_{i,t-1} + \alpha_i + \varepsilon_{it}$, with ε_{it} i.i.d. normal $(0,1)$. We assume one-dimensional underlying heterogeneity: $\alpha_i = \alpha(\xi_i)$, $\rho_i = \rho(\xi_i)$, ξ_i scalar. In addition, $|\rho_i| < 1$. We use $h_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ in the first estimation step. In this case $\mathrm{plim}_{T\to\infty}\, h_i = \frac{\alpha(\xi_i)}{1-\rho(\xi_i)}$. We focus on the average effect $M = \frac{1}{N}\sum_{i=1}^{N} \rho_i$. The specification we use is as follows: $\xi_i \sim \mathcal{N}(0,1)$, $\alpha_i = \xi_i$, and $\rho_i = \frac{\exp(.5\alpha_i+1)-1}{\exp(.5\alpha_i+1)+1}$. We also assume a stationary initial condition Y_{i1}. Figure C1 shows the results for fixed-effects and two-step grouped fixed-effects estimators, for $K = 5$ and $K = 10$.

Figure C1: Heterogeneous AR(1) model ($N = 500$, $T = 10$, 200 replications)



Notes: Monte Carlo distributions. Blue is fixed-effects, green is two-step grouped fixed-effects. Bottom graphs obtained using split-panel jackknife bias reduction. $M_0 \approx .44$.

C.2 Double grouped fixed-effects in a probit model

An alternative estimator, in linear or index models, is "double" grouped fixed-effects. As an example, consider the static probit model

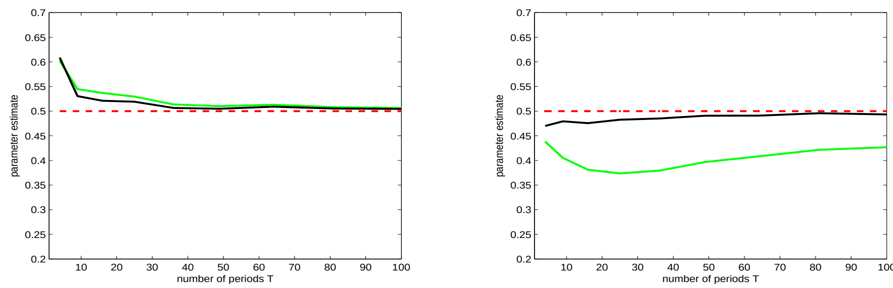
$$Y_{it} = \mathbf{1}\{X_{it}'\theta + \alpha_i + \varepsilon_{it} \geq 0\},$$

where ε_{it} are i.i.d standard normal, and $X_{it} = \mu_i + v_{it}$, v_{it} i.i.d, independent of $\mu_i, \alpha_i, \varepsilon_{is}$.

Consider the moments $h_i = (\overline{Y}_i, \overline{X}_i')'$. In the first step, we discretize each component of h_i separately. In the second step, we estimate the probit model while including all group indicators additively. Figure (C2) shows the results in two cases: one-dimensional (panel A) and two-dimensional heterogeneity (panel B). The results compare two-step grouped fixed-effects with "double" grouped fixed-effects estimators.

Figure C2: Static probit model ($N = 100$, 100 replications)

A. $\alpha_i = \mu_i$ B. $\text{Corr}(\alpha_i, \mu_i) = .5$



Notes: Averages over simulations. Green is two-step grouped fixed-effects, black is double grouped fixed-effects, red is true value. $K = \left\lfloor \sqrt{T} \right\rfloor$.

C.3 High-dimensional fixed-effects

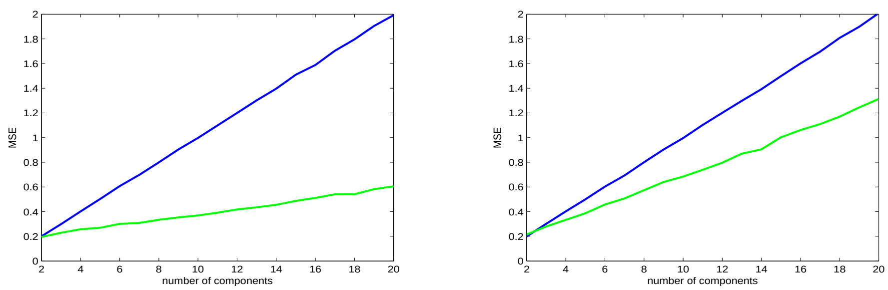
Let $Y_{ijt} = \alpha_{ij} + \varepsilon_{ijt}$, $\alpha_{ij} = \alpha_j(\xi_i)$, where $j = 1, ..., q$. We focus on the MSE:

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{q} \left(\widehat{\alpha}_j(\widehat{k}_i) - \alpha_{ij} \right)^2 .$$

We use the following specification: ε_{ijt} i.i.d standard normal, and either (case A) $\alpha_j(\xi_i) = \xi_i$ standard normal, or (case B) $\alpha_j(\xi_i) = \xi_{i1} + \xi_{i2}\Phi^{-1}\left(\frac{j}{q+1}\right)$, where Φ is the standard normal cdf; ξ_{i1} standard normal, ξ_{i2} independent log-normal. Figure C3 shows the results for fixed-effects and two-step grouped fixed-effects estimators, for $K = 10$.

Figure C3: High-dimensional fixed-effects ($N = 100$, $T = 10$, $K = 10$, 100 replications)

A. One-dimensional B. Two-dimensional



Notes: Mean squared errors. Blue is fixed-effects, green is two-step grouped fixed-effects.

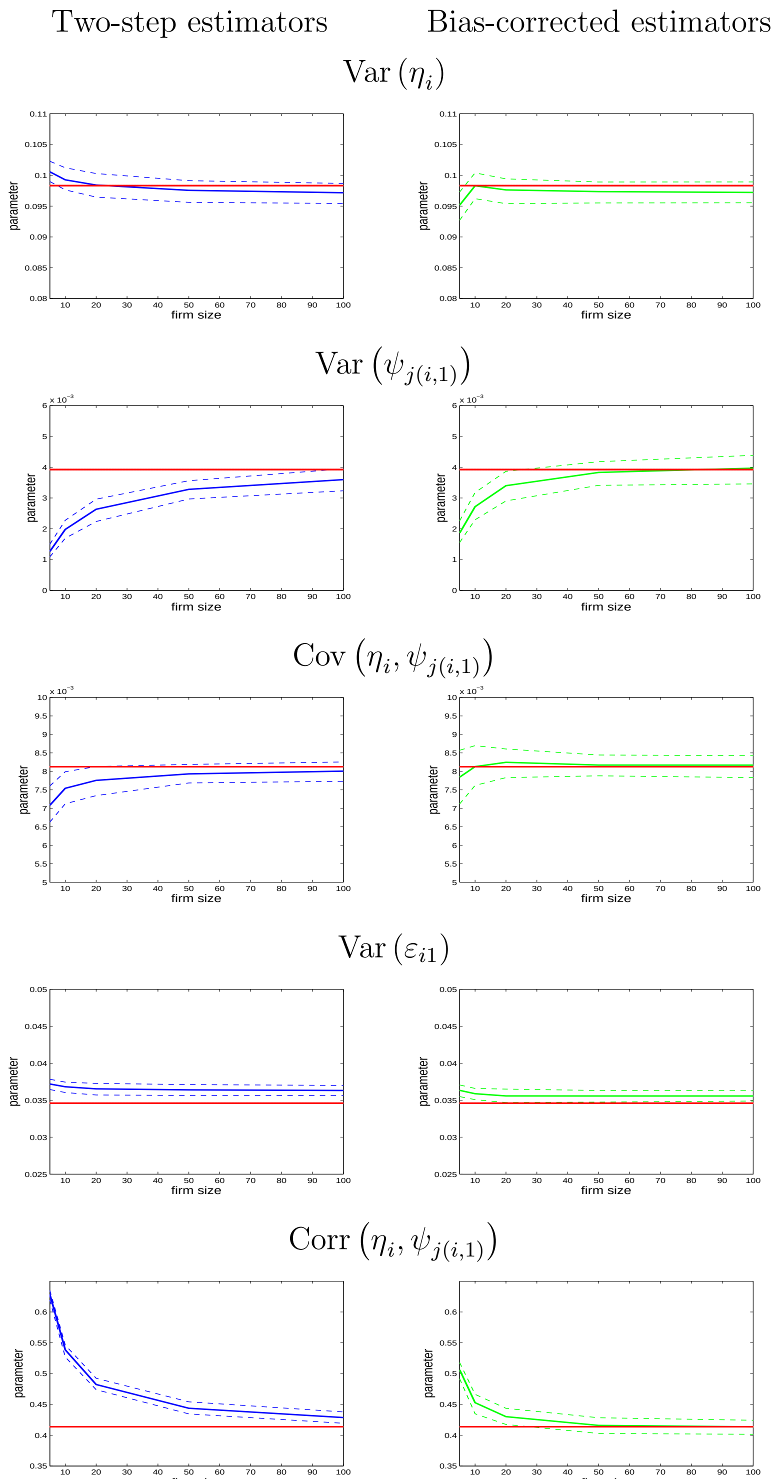
D Additional results

Table D1: Firm and worker effects, discrete data generating process

	$\mathrm{Var}(\eta_i)$	$\mathrm{Var}\left(\psi_{j(i,1)}\right)$	$\mathrm{Cov}\left(\eta_i, \psi_{j(i,1)}\right)$	$\mathrm{Var}(\varepsilon_{i1})$	$\mathrm{Corr}\left(\eta_i, \psi_{j(i,1)}\right)$
	True values				
	.0983	.0039	.0081	.0346	.4137
	Two-step estimator, means				
Firm size = 5	.1006	.0013	.0071	.0372	.6286
Firm size = 10	.0993	.0020	.0075	.0368	.5386
Firm size = 20	.0984	.0026	.0078	.0365	.4823
Firm size = 50	.0976	.0033	.0079	.0364	.4437
Firm size = 100	.0972	.0036	.0080	.0363	.4287
	Two-step estimator, standard deviations				
Firm size = 5	.00087	.00011	.00027	.00036	.00413
Firm size = 10	.00094	.00015	.00022	.00037	.00512
Firm size = 20	.00101	.00019	.00020	.00043	.00485
Firm size = 50	.00081	.00018	.00014	.00041	.00532
Firm size = 100	.00079	.00020	.00014	.00041	.00508
	Bias-corrected estimator, means				
Firm size = 5	.0952	.0019	.0078	.0363	.5061
Firm size = 10	.0983	.0027	.0081	.0359	.4528
Firm size = 20	.0976	.0034	.0083	.0356	.4301
Firm size = 50	.0974	.0038	.0082	.0356	.4158
Firm size = 100	.0972	.0040	.0082	.0356	.4135
	Bias-corrected estimator, standard deviations				
Firm size = 5	.00121	.00018	.00038	.00042	.00742
Firm size = 10	.00108	.00023	.00028	.00041	.00816
Firm size = 20	.00105	.00026	.00023	.00047	.00682
Firm size = 50	.00085	.00023	.00016	.00044	.00687
Firm size = 100	.00080	.00024	.00016	.00042	.00632

Notes: Mean and standard deviation of parameters across simulations. $K = 10$. 100 *simulations.*

Figure D4: Firm and worker heterogeneity, discrete DGP



Notes: Means (solid) and 95% confidence bands (dashed). Blue is two-step, green is bias-corrected estimator. The red curve indicates the true parameter value. 100 *replications.*